BOOTSTRAP INFERENCE FOR PARAMETRIC QUANTILE REGRESSION

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2011

Tatjana Kecojević
School of Mathematics

Contents

Word count 86139

List of Tables

List of Figures

The University of Manchester

Tatjana Kecojević
Doctor of Philosophy
Bootstrap Inference for Parametric Quantile Regression
November 19, 2011

The motivation for this thesis came from the provision of a large data set from Saudi Arabia giving anthropometric measurements of children and adolescents from birth to eighteen years of age, with a requirement to construct growth charts. The construction of these growth charts revealed a number of issues particularly in the respect to statistical inference relating to quantile regression.

To investigate a range of different statistical inference procedures in parametric quantile regression in particular the estimation of the confidence limits of the τ^{th} ($\tau \in [0, 1]$) quantile, a number of sets of simulated data in which various error structures are imposed including homoscedastic and heteroscedastic structures were developed. Methods from the statistical literature were then compared with a method proposed within this thesis based on the idea of Silverman's (1986) kernel smoothing. This proposed bootstrapping method requires the estimation of the conditional variance function of the fitted quantile. The performance of a variety of variance estimation methods combined within the proposed bootstrapping procedure are assessed under various data structures in order to examine the performance of the proposed bootstrapping approach. The validity of the proposed bootstrapping method is then illustrated using the Saudi Arabian anthropometric data.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I would like to thank my supervisor, Dr. Peter Foster, for his advice over the years.

Special thanks to my fellow student, Therese Sheppard and Andreas Vrahimis, for their friendship, endless hours of discussions of our works and good humour.

Special thanks are reserved for my colleagues and friends at UCLAN, Ian McGowan and George Rawlings, whose friendship, support and encouragement have enabled me to do this research.

I wish to express my deep gratitude to my parents, Katarina and Milan, to whom this thesis is dedicated, and to my sister Ivana for their love and encouragement in persuing my goals as long as I can remember.

Last, but by no means least, my huge appreciation and thanks go to Alan for his love, understanding and great patience at all times.

Publication

P.J. Foster, T. Kecojević. Construction of the Reference Growth Charts for Saudi Arabian Children and Adolescents. *Advances in Social Science Research Using R; Series: Lecture Notes in Statistics; Subseries: Lecture Notes in Statistics Proceedings*; Springer-Verlag, New York, 196:107-128, 2010.

Chapter 1

Introduction

The purpose of regression analysis is to explore the relationship between a stochastic response variable and a number of explanatory variables, some of which may be stochastic. Traditional regression analysis focuses on the estimation of the means of a response variable conditional on the values of the explanatory variables. The idea of modeling and fitting the *conditional mean function* is at the heart of the regression modeling techniques. Under ideal conditions of normality assumption and homscedasticity, they are capable of providing a complete and parsimonious description of how the mean of response variable depends on the values of the conditioning covariate variables. However, the mean is only one of the characteristics of the conditional distribution that is of interest. A much fuller description of the conditional dependence could provide the information about scale and shape of the response distribution. This fuller and richer description of the conditional distribution is exploited to allow the estimation of the entire distribution of the response variable conditional upon a set of explanatory variables. In quantile regression the conditional mean is replaced by a defined set of conditional quantiles which are able to provide a more complete view of the underlying relationships between the response variable and the explanatory variables.

The idea of quantile regression goes back to the early work on conditional-median regression modeling by Ruđer Josip Bosković in 1760's (Koenker, 2005). However,

it was not until late 1970's that median regression modeling, powered by the developments in the computational science, become a practical statistical tool that can be easily implemented by the application of the linear programming optimaization algorithm of minimising a generalised measure of distance.

Over the last four decades, quantile regression model has received significant attention in the theoretical literature. Koenker and Bassett (1978) demonstrated how quantile regression models can be easily estimated. Statistical software for quantile regression is widely available in many well known statistical packages. Koenker has developed `quantreg` package in `R` (public domain language for data analysis, 2011) which enables easy application of the method. In Chapter 2 we review some basic theoretical principles of the linear quantile model and present fundamental asymptotic approximation theory and inferential strategies (Bassett and Koenker, 1978).

Quantile regression has been used in a wide range of application settings. *Growth charts* (also known as centile reference charts) are developed using quantile regression, which are widely used in medical science. *Reference growth curves* for childrens' anthropometric measurements have a long history in medicine. The first study of this kind can be found as early as XIX century, conceived by the Belgian social statistician Adolphe Quetelet (Wei et al., 2006). The nutritional assessment and diagnosis of over and under nourished children have financial application, and it is now globally acknowledged that investment in human resources is a pre-request for economic developments of any nation (WHO, 2009).

The conventional method of constructing reference growth centiles within a growth chart is to obtain empirical percentiles from cross-sectional data at a particular set of time points, and to fit a smooth curve to them (Wei and He, 2006). In recent years several other methods have been developed, but the most popular in medical circles is the *lambda, mu, sigma* (*LMS*) method proposed by Cole (1988). The LMS method assumes that at any time point t, a physical measurement $Y(t)$ can be transformed to the approximately standardised normal $Z(t)$ using a Box-Cox transformation (Box and Cox, 1964). The distribution of the measurement is summarised by three parameters, the Box-Cox power λ (L), the mean μ (M) and the coefficient of variation σ (S),

giving rise to the name LMS. Cole and Green (1992) proposed a penalised likelihood approach to smooth the resulting LMS percentile curves and the methodology of this method is described in Capter 8.

In Chaper 8 the application of the LMS method is illustrated using the Saudi Arabian data, which was collected as a part of a nationwide project to establish normal anthropometric measurements for Saudi Arabian children and adolescents. The reference values were derived from cross-sectional data applying the LMS method of Cole and Green (1992) using the `lmsqreg` package in `R`, which was developed by Carey (2002). This chapter provides an overview of how the method has been applied, more specifically how the relevant issues concerning the construction of the growth charts have been addressed and is illustrated by just using the girls weight data (birth to three years old). These issues include identifying the outliers, diagnosing the appropriate amounts of smoothing and averaging the reference standards for the overlapping 2 to 3 year age range. The use of ANCOVA has been introduced and illustrated as a tool for making growth standard comparisons between different geographical regions and between genders.

Assessing the accuracy of the τ^{th} ($\tau \in [0, 1]$) quantile parametric regression function estimate requires valid and reliable procedures for estimating the asymptotic variance-covariance matrix of the estimated parameters. This covariance matrix depends on the reciprocal of the density function of the error (sparsity function) evaluated at the quantile of interest which, particularly for non-iid cases, results in a complex and difficult estimation problem. It is well-known that the construction of the confidence intervals based on the quantile regression estimator can be simplified by using a bootstrap. In the iid case various bootstrap procedures can be effectively used. However, in the non-iid case use of the bootstrap is much more difficult. A number of proposals have been made, which includes Parzen, Wei and Ying's method based on pivotal estimating functions (Parzen et al., 1994) and the Markov Chain Marginal (Kocherginsky et al., 2005) bootstrap method. In Chapter 3 we propose an alternative bootstrap method based on the idea of Silverman's (1986) kernel smoothing approach. After fitting τ^{th} quantile function, we obtain the residuals u_i which

being squared and centered to zero su_i are assumed to follow a gamma distribution. Estimating the mean function of the centered squared residuals gives the conditional variance function of the error term of the τ^{th} quantile. Using this estimate of the conditional variance function allows the standardisation of the residuals which are used in Silverman's (1986) kernel smoothing bootstrapping procedure to obtain the bootstrap estimate of τ^{th} quantile function's parameters.

By modeling the conditional mean of the squared residuals using gamma GLM, we are able to estimate the conditional variance function of the τ^{th} quantile under the assumption of the error being normally distributed. In Chapter 4 we explore the robustness of this method when the error is not normal and therefore the residuals are non gamma. Furthermore, in practice we are also rarely confident that constant coefficient of variation (CV) assumption of the gamma distribution is true and therefore we examine different approaches that would allow the CV not to be constant.

In Chapter 5 we investigate the problem of non-homogenious dispersion in GLM models by the options provided by *Double Generalised Linear Models*, which models the mean and dispersion simultaneously in a context of general linear model, allowing for heteroscedastic dispersion (Smyth and Verbyla, 1999a). We also relax the assumption that the su_i's are strictly gamma distributed by assuming that they are from a *Tweedie* family of distributions (Smyth and Jorgensen, 1999) and we investigate the flexibilities offered by the *Quasi-Likelihood* GLMs.

Chapters 6 and 7 look into the possibilities of modeling the conditional variance function by applying non-parametric approaches. In Chapter 6 we consider the approach based on smoothing the squared residuals by fitting a local polynomial, whilst in Chapter 7 we examine the use of the difference-based approach. The residuals obtained from the estimated quantile regression function are used to calculate the differences that will act as the response variable in the conditional variance model estimation. In the case of the difference based variance estimation the squared pseudoresiduals of order k, $\Delta^2_{k,i}$ construct a k-dependent correlated sequence, a problem for the application of the cross-validation approach that is required for the estimation of the bandwidth parameter h. This thesis proposes the use of the leave-one-out+$2k$

cross-validation method to overcome the correlation problem in the k-dependent correlated sequence.

Chapter 2

Quantile Regression

2.1 Introduction

After almost four decades of development *quantile regression* is gradually emerging as a fundamental tool of applied statistics. Complementing the use of *least squares methods* for estimating conditional mean models, quantile regression not only offers a more robust alternative for estimating the central tendency of the response, but also allows more detailed exploration of the conditional distribution of the response.

In classical regresion framework, the conditional mean $E(Y|X = x)$ is estimated by the least squares method. Unless, one is willing to make more strong distributional assumption on the conditional distribution of the response, least square method does not provide any information beyond the conditional mean. Quantile regression generalises the idea of median regression to estimation of other conditional quantile functions by minimizing weighted sums of absolute residuals and enables examination of the effect of various covariates on the quantile of the response. By supplementing the estimation of the conditional mean functions, quantile regression is more effective in enabling a complete statistical analysis of the stochastic relationships between the random variables.

2.2 Quatiles and Quantile Functions

We use the cumulative distribution function, *cdf*, to describe the distribution of a random variable Y. By using the *cdf* function F we can get for every value of y the proportion of the population for which $Y \leqslant y$

$$F(y) = P(Y \leqslant y).$$

Hence, for any range of y, we can calculate the the proportion of the population. Having a *cdf* of a random variable Y we can define the *quantile function*

$$\mathcal{Q}(\tau) = F^{-1}(\tau) = inf\{y|F(y) \geqslant \tau\}, \qquad 0 < \tau < 1.$$

Figure 2.1 shows the cdf of the standard normal distribution on which we can identify $F(0) = 0.5$ and $F(0.95) = 1.644854$, which are called the τ^{th} quantiles of Y: 0.5^{th} and 0.95^{th} quantile respectively. The 0.5^{th} quantile is what we call the median. Equivalently, for a sample $y_1, y_2, \ldots, y_n$ we use the *emirical cdf*, $\hat{F}$, which gives the proportion of the sample values that is less than or equal to any given y, for which we can also define the *empirical quantile function* as $\hat{\mathcal{Q}}(\tau) = \hat{F}^{-1}(\tau)$.

Observing quantiles and quantile functions leads to a fuller collection of the summary measurements of the random variable Y. The quantiles are related to order statistics as the solution to a relatively simple optimisation problem. Koenker and Hallock (2001) suggest that as an alternative to finding the mean as the solution to the problem of minimizing the sum of squared residuals, the median can be found as the solution to the minimization of a sum of absolute residuals. By further applying this idea other quantiles can be estimated.

To illustrate this, let us define the linear, peace-wise function

$$\rho_\tau(u) = u\big(\tau - I(u < 0)\big), \qquad 0 < \tau < 1, \tag{2.1}$$

where

$$I(u < 0) = 1 \quad if \quad u < 0, \quad \text{and } 0 \text{ otherwise.} \tag{2.2}$$

The example of $\rho_{\tau=0.75}(u)$ is illustrated in Figure 2.2. The τ^{th} sample quantile of Y

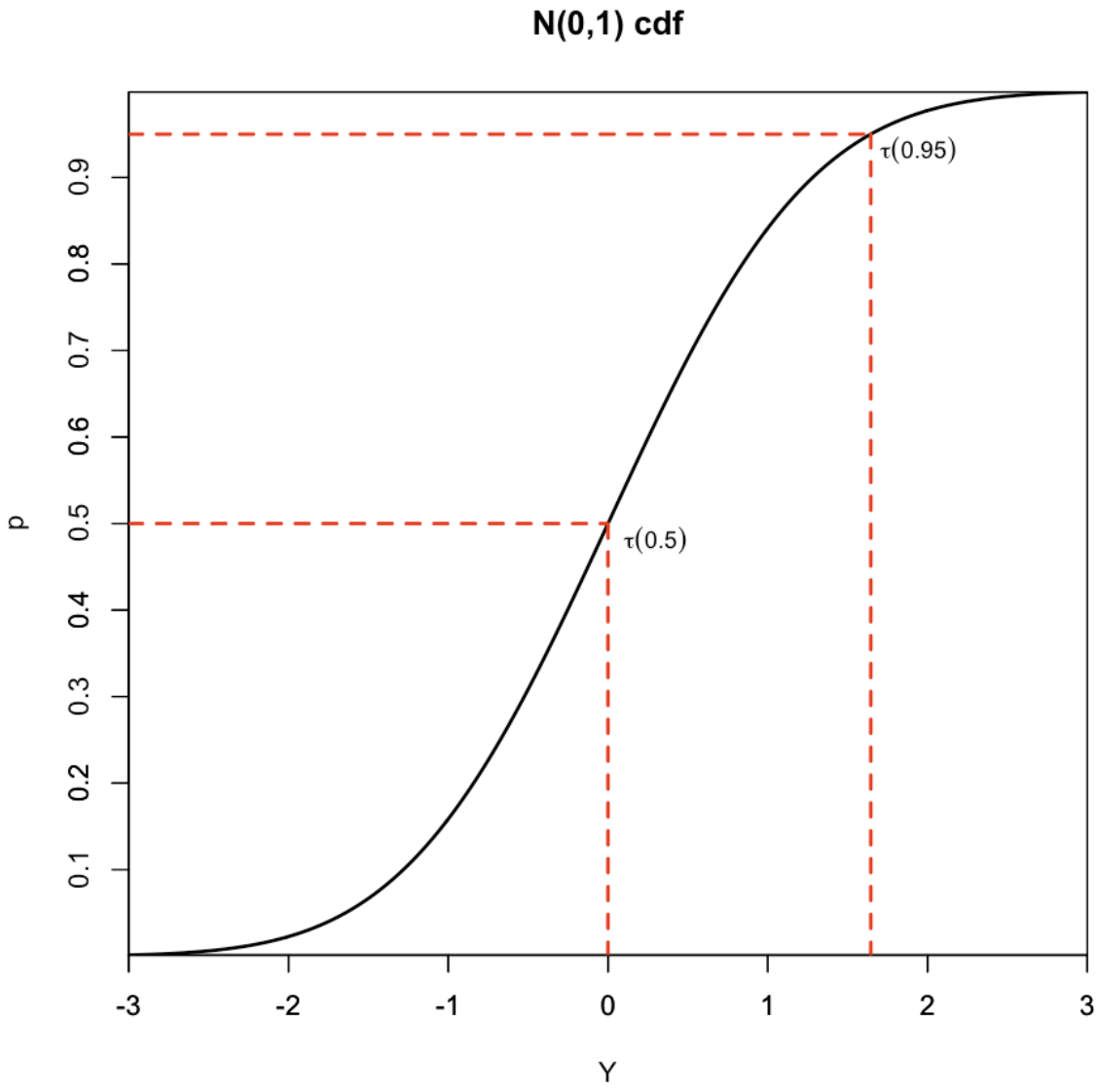


Figure 2.1: *CDF. The cdf of the standard normal distribution.*

is a solution, $\hat{\xi}$ to

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi), \tag{2.3}$$

which can be found by minimizing the expectation of $\rho_\tau(Y - \xi)$ with respect to ξ. This yields solutions, $\hat{\xi}(\tau)$, the smallest of which is the τ^{th} quantile. The minimum value $\hat{\xi}$ of

$$R(\xi) = \sum_{i=1}^{n} \rho_\tau(y_i - \xi), \tag{2.4}$$

is found when the value of $R(\xi)$ increases for either positive or negative changes in ξ

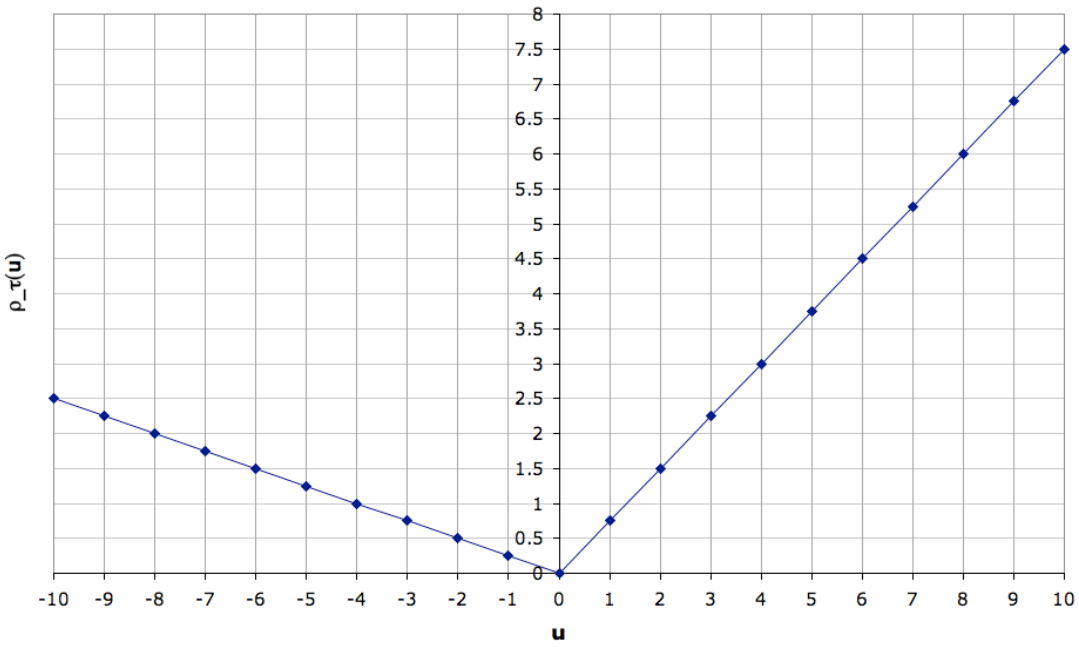


Figure 2.2: *Sample quantile ρ function.*

around $\hat{\xi}$ (Koenker, 2005, Chapter 1).

The optimization problem of defining the 0.25^{th} quantile of a small random sample

$$y_i : \{1.62234, -0.16122, -0.46060, -0.61184, 0.67386, 0.71067, -0.86516\},$$

of size $n = 7$, from the standard normal distribution, is illustrated in Figure 2.3. Let us guess and assume that the fifth sample element corresponds to the 0.25^{th} sample quantile, $\hat{\xi} = 0.67386$. For each y_i we calculated the value of the loss function

$$\rho_{\tau=0.25}(u_i) : \{0.23712, 0.62631, 0.85085, 0.96428, 0.00000, 0.00920, 1.15427\}$$

and after adding them up it provides the value of the objective function $R(\xi) = 3.8420175$. Minimising $R(\xi)$ with respect to ξ yields the optimal solution $\hat{\xi} = -0.61184$. If we again calculated the values of the loss function for the optimal $\hat{\xi}$, we get

$$\rho_{\tau=0.25}(u_i) : \{0.55855, 0.11266, 0.03781, 0.00000, 0.32143, 0.33063, 0.18999\}$$

which gives the minimum value for $R(\xi)$ of 1.55105. If we would like to identify the corresponding 0.75^{th} quantile of the given sample we would get the minimum

value the objective function $R(\xi)$ to be 1.92833 for $\hat{\xi} = 0.71067$ (see Figure 2.3). By sorting the sample in the order it is easy to identify those two sample elements that correspond to the $\tau = 0.25$ and $\tau = 0.75$

$$y_i : \{-0.86516, \mathbf{-0.61184}, -0.46060, -0,16122, 0.67386, \mathbf{0.71067}, 1.62234\}.$$

As we can see, it is easy to define the sample quantile in terms of order statistics, $y(1) \leqslant y(2) \leqslant \cdots \leqslant y(n)$, by making a sorting rearrangements of the original sample. Nevertheless, their formulation as a minimisation problem offers the advantage that provides a natural generalisation of the quantiles to the quantile regression.

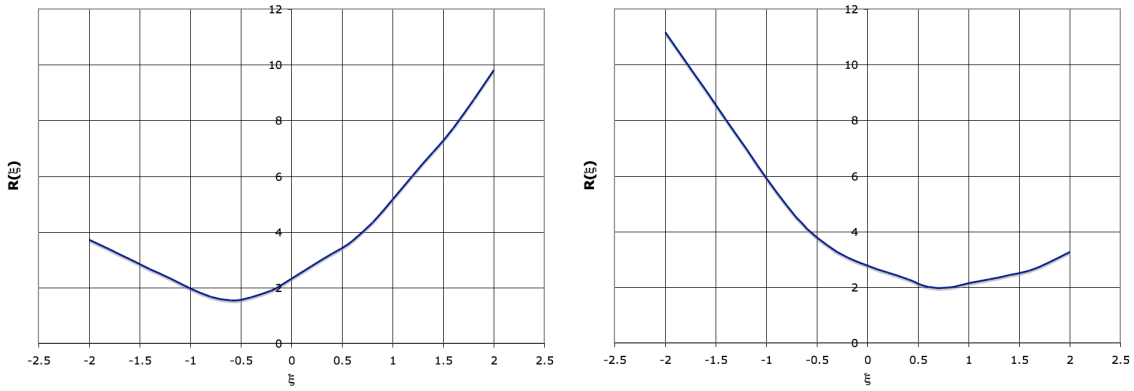


Figure 2.3: *The objective function $R(\xi)$. Figures illustrates the objective function for the optimization problem of defining the 0.25^{th} quantile (left panel) and 0.75^{th} quantile (right panel) of a small random sample of size $n = 7$, from the standard normal distribution.*

Following this discussion of sample quantile, we can find conditional quantiles. The familiar least square regression model provides relatively simple framework within which to discuss the ideas of conditional quantiles.

For a random sample $\{y_1, y_2, \ldots, y_n\}$, the sample mean is a solution of the problem

$$\hat{\mu} = \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2, \tag{2.5}$$

which represents an estimate of the unconditional population mean, $E(Y)$. Replacing the scalar μ by a conditional mean of y for a given $\boldsymbol{x}$, which we write as a linear function of the unknown parameters $\boldsymbol{\beta}$, $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$, and by solving

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}^\top \boldsymbol{\beta})^2, \tag{2.6}$$

we obtain the estimate of the conditional mean function $E(Y|\boldsymbol{x})$. Thus, to get an estimate of the conditional median function we replace the scalar ξ in (2.3) by a parametric function $\xi(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}$ and set $\tau = 1/2$. Figure 2.4 plots data from

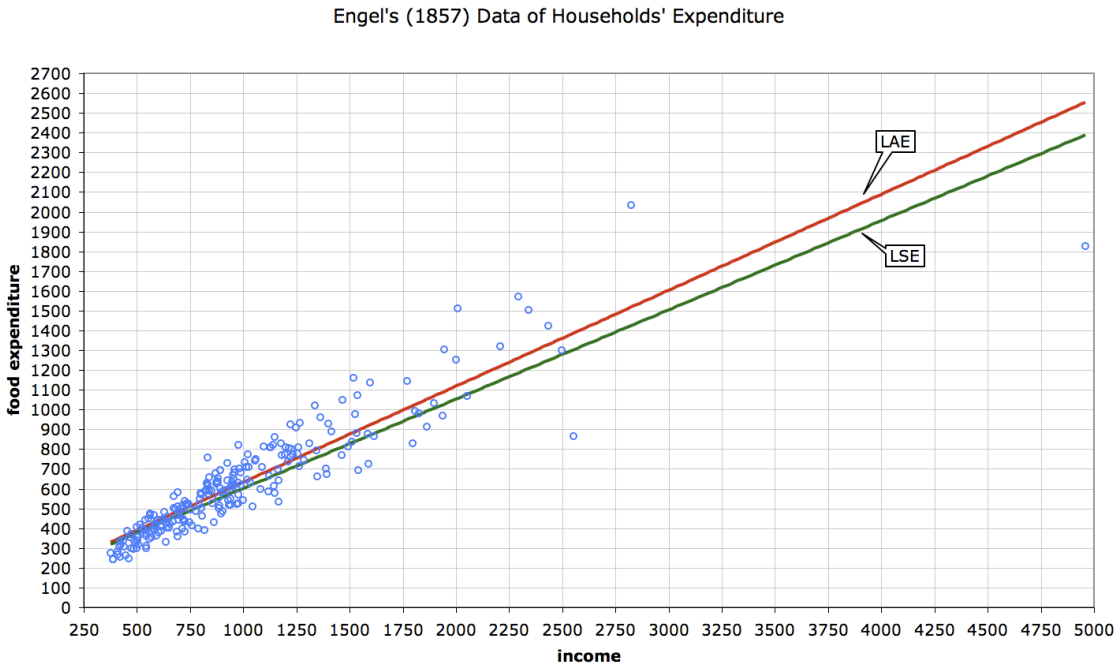


Figure 2.4: *Eengel's (1857) Data: Figure shows a scatterplot of household's expenditure on food versus annual income. The data consists of 235 observations on European working class house holds. Superimposed are east squared estimate (LSE: green line) and least absolute residual estimate (LAE: red line).*

Engel's (1857) study of households' expenditure on food versus annual income. The data is available in `R` in the package `quantreg`. Superimposed are the regression mean function obtained by least squared estimation (LSE), green line, and estimate of the median regression function obtained by minimizing the sum of absolute residuals (LAE), red line.

Any other conditional quantile function $\mathcal{Q}_y(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}(\tau)$ can be estimated by solving

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}^\top\boldsymbol{\beta}). \tag{2.7}$$

Figure 2.10 illustrates several quantile function estimates for a bivariate data (weight depending on age).

2.3 Distribution of a Univariate Quantile

In the standard linear regression model inference depends on the assumption that the errors are independently, identically and normally distributed (iind). Within a quantile regression context it would be desirable to provide equivalent finite sample and procedure as exist in many regression model testing methods. However, it is almost always the case that such finite sample results are intractable and the best that can be hoped for is the establishment of some large sample asymptotic results. Having said that, quantile regression asymptotic theory can be derived by examining the inference results for the univariate quantile. This theory is well acknowledged with many of the important results developed in Koenker (2005). Theorems from various sources are given and discussed in this chapter by way of providing a literature review of key results for quantile regression.

2.3.1 Quantile Regression Asymptotic

Let $\{x_i\colon\ i = 1, \ldots,\ n\}$ denote rows of a known $(n \times k)$ design matrix and suppose $\{y_i\colon\ i = 1, \ldots,\ n\}$ is a random sample on the regression process $u_i = y_i - x_i\beta$ which has distribution function F as its *cdf*. We can define the τ^{th} regression quantile, $0 < \tau < 1$, as any solution to:

$$\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \tag{2.8}$$

and by referring to the results of the clasical theory of inference for the univariate quantile we can derive the finite-sample density of the regression quantile estimation, $\hat{\beta}(\tau)$, in a similar manner to the derivation of the density in the one-sample case.

Let us redefine function g_n that take into account the sample size n

$$g_n\left(\xi_\tau\right) = \frac{1}{n} \sum_{i=1}^{n} \left(I\left(Y_i\ <\ \xi_\tau\right) - \tau\right). \tag{2.9}$$

By examining Figure 2.3, $g_n(\xi)$ is evidently monotonically increasing in ξ and, by monotonicity, $\hat{\xi}_\tau$ is greater than ξ if and only if $g_n\left(\xi\right)\ <\ 0$. By developing this finding appropriate distributional characteristics, for example the mean and the variance, of

$g_n(\xi_\tau)$ can be found (Koenker, 2005, Chapter 3):

$$E\left[g_n\left(\xi_\tau + \delta/\sqrt{n}\right)\right] = \left(F\left(\xi_\tau + \delta/\sqrt{n}\right) - \tau\right) \to f\left(\xi_\tau\right)\delta/\sqrt{n}$$

and

$$V\left[g_n\left(\xi_\tau + \delta/\sqrt{n}\right)\right] = F\left(\xi_\tau + \delta/\sqrt{n}\right)\left(1 - F\left(\xi_\tau + \delta/\sqrt{n}\right)\right)/n \to \tau\left(1-\tau\right)/n.$$

This approach has been further developed and presented in the following:

Theorem 2.3.1. *(Koenker and Bassett (1978, Thm.4.1)) Let* $\left\{\hat{\xi}_{n\tau_1}, ...\hat{\xi}_{n\tau_m}\right\}$ *with* $0 < \tau_1 < \tau_2 < ... < \tau_m < 1$*, denote a sequence of unique sample quantiles based on a random sample of size n from a population with inverse distribution function* $\xi_\tau = F^{-1}(\tau)$*. If F is continuous and has continuous and positive density, f, at* ξ_{τ_i}*,* $i = 1, ... m$*, then,*

$$\sqrt{n}\left[\hat{\xi}_{n\tau_1} - \xi_{\tau_1}, ... \hat{\xi}_{n\tau_m} - \xi_{\tau_m}\right]$$

converges in distribution to an (m)*-variate Gaussian random vectors with mean,* 0*, and covariance matrix* $\mathbf{\Omega}\left(\tau_1, ... \tau_m;\ F\right)$ *with typical element*

$$\omega_{ij} = \frac{\tau_i\left(1-\tau_j\right)}{f\left(\xi_{\tau_i}\right) f\left(\xi_{\tau_j}\right)}, \quad i \leqslant j.$$

This means that we can set $\omega = \tau(1-\tau)/f^2(\xi_\tau)$, and express

$$\begin{aligned} P\left\{\sqrt{n}\left(\hat{\xi}_\tau - \xi_\tau\right) > \delta\right\} &= P\left\{\frac{g_n\left(\xi_\tau + \delta/\sqrt{n}\right) - f\left(\xi_\tau\right)\delta/\sqrt{n}}{\sqrt{\tau\left(1-\tau\right)/n}} < -\omega^{-1}\delta\right\} \\ &\to 1 - \Phi\left(\omega^{-1}\delta\right). \end{aligned}$$

and hence conclude that

$$\sqrt{n}\left(\hat{\xi}_\tau - \xi_\tau\right) \rightsquigarrow \mathcal{N}\left(0, \omega^2\right). \tag{2.10}$$

There is an obvious resemblance in the asymptotic behavior between sample quantiles in the location model and regression quantiles in the linear models. In another theorem Koenker and Bassett (1978), makes this similarity even more explicit:

Theorem 2.3.2. *(Koenker and Bassett (1978, Thm.4.2)) Let* $\left\{\hat{\boldsymbol{\beta}}_n\left(\tau_1\right), \hat{\boldsymbol{\beta}}_n\left(\tau_2\right), ... \hat{\boldsymbol{\beta}}_n\left(\tau_m\right)\right\}$ *with* $0 < \tau_1 < \tau_2 < ... < \tau_m < 1$*, denote a sequence of unique sample quantiles based on a random sample of size n from a population with inverse distribution function* $\xi_\tau = F^{-1}(\tau)$*,* $\boldsymbol{\xi}_\tau = (\xi_\tau, 0, ... 0) \in \mathbb{R}^k$ *and* $\hat{\boldsymbol{\xi}}_{n\tau} = \hat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}$*. Assume:*

1. *F is continuous and has continuous and positive density, f, at ξ_{τ_j}, $j = 1, ... m$, and*

2. *$x_{ji} = 1 : i = 1, 2, ...n$ and $\lim_{n\to\infty} n^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{Q}_0$, a positive definite matrix.*

Then,

$$\sqrt{n}\left[\hat{\boldsymbol{\xi}}_{n\tau_1} - \boldsymbol{\xi}_{\tau_1}, ..., \hat{\boldsymbol{\xi}}_{n\tau_m} - \boldsymbol{\xi}_{\tau_m}\right]$$

converges in distribution to an (mk)-variate Gaussian random vector with mean 0 and covariance matrix $\boldsymbol{\Omega}(\tau_1, ..., \tau_m; F) \otimes \boldsymbol{Q}_0^{-1}$, where $\boldsymbol{\Omega}$ is the covariance matrix of the coresponding m ordinary sample quantiles obtained from a random sample of size n from distribution F.

Applying these theorems the joint distribution of several quantiles can be obtained, $\hat{\boldsymbol{\zeta}}_n = \left(\hat{\xi}_{\tau_1}, ..., \hat{\xi}_{\tau_m}\right)$ with $\boldsymbol{\zeta}_n = (\xi_{\tau_1}, ..., \xi_{\tau_m})$,

$$\sqrt{n}\left(\hat{\boldsymbol{\zeta}}_n - \boldsymbol{\zeta}\right) \rightsquigarrow \mathcal{N}(0, \boldsymbol{\Omega}) \tag{2.11}$$

where $\boldsymbol{\Omega}$ is an $m \times m$ matrix with the following elements:

$$(\omega_{ij}) = \frac{(\tau_i \wedge \tau_j - \tau_i\tau_j)}{(f(F^{-1}(\tau_i))\, f(F^{-1}(\tau_j)))}$$

in which

$$(\tau_i \wedge \tau_j - \tau_i\tau_j) = \begin{cases} \tau_i(1-\tau_i) & i = j \\ \tau_i(1-\tau_j) & i < j \\ \tau_j(1-\tau_i) & i > j. \end{cases}$$

2.3.2 Quantile Regression Asymptotics with IID Errors

The results for the classical regression model

$$y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + e_i \tag{2.12}$$

where $\{e_i\}$ is iid, can be derived by analogy with the results from the ordinary one sample quantile model.

Portnoy (2010) derived the key asymptotic distributional results for the mp-variate quantle regression estimators

$$\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n\left(\tau_j\right)-\boldsymbol{\beta}_n\left(\tau_j\right)\right)\right)_{j=1}^{m} \rightsquigarrow \boldsymbol{\mathcal{N}}_{m\times p}\left(0,\left(\boldsymbol{A}_m^{-1}\boldsymbol{\Omega}_m\boldsymbol{A}_m^{-1}\right)\otimes\boldsymbol{Q}_0^{-1}\right) \tag{2.13}$$

where $\boldsymbol{A}_m$ is the $m\times m$ diagonal matrix with diagonal elements $f\left(F^{-1}\left(\tau_j\right)\right)$ and $\boldsymbol{\Omega}_m \equiv \left(\omega_{ij}\right)_{m\times m}$; $\omega_{ij} \equiv \left(\tau_i \wedge \tau_j - \tau_i\tau_j\right)$.

In this case, when the errors have precisely the same distribution whatever values may be taken by the components of the vector $\boldsymbol{x}_i$, we refer to this as the *pure location shift model*. It assumes that $\boldsymbol{x}_i$ does not affect any other aspects of the conditional distributional shape of the response y apart from its location.

The form of $\boldsymbol{\beta}\left(\tau_j\right)$ is relatively simple in the case of iid regression model. We construct a pure location shift model in which the conditional quantiles are parallel. Assuming that the first component of $\boldsymbol{\beta}$ corresponds to the intercept, we have $\boldsymbol{\beta}\left(\tau\right) = \boldsymbol{\beta}+\xi_\tau\boldsymbol{\epsilon}_1$, where $\xi_\tau = F^{-1}\left(\tau\right)$ and $\boldsymbol{\epsilon}_1$ is the $(p\times 1)$ first unit basis vector $(1,0,\ldots,0)^\top$ of $\mathbb{R}^p$ and where the covariance matrix $\boldsymbol{\Omega}$ has the same form as in the one-sample setting. This outcome enables a Wald-type statistic to be used for inference in the iid quantile regression modelling (Koenker, 2005, Chapter 3).

2.3.3 Quantile Regression Asymptotics with Non-IID Errors

The least squares regression model provides an examination of the central tendency of the response variable, under the strong assumption that the errors are iind, which in practice is rarely the case. Nevertheless, the quantile regression model offers a full picture of the distribution of the response variable which may be influenced by the covariate in many other ways: "expanding its dispersion as in clasical models of heteroscedasticity, stretching one and compresing the other tail of the distribution, and even including multimodality" (Koenker and Hallock, 2001). In this way the quantile regression model offers a fuller description of the relationship between the response variable and the covariate.

In non-iid error settings, the asymptotic theory of $\hat{\boldsymbol{\beta}}\left(\tau\right)$ is relatively more complex than that in the iid error, location shift model. Huber (1964) gave an appropriate

covariance matrix of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right)$ of the following form:

$$\tau(1-\tau)\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}, \tag{2.14}$$

where

$$\boldsymbol{J}(\tau) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\top} \tag{2.15}$$

and

$$\boldsymbol{H}(\tau) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\top} f_i\left(\xi_i(\tau)\right). \tag{2.16}$$

The conditional density of the response, y_i, evaluated at the τ^{th} conditional quantile is given by $f_i(\xi_{i\tau})$. As we have seen earlier, in the iid cases the f_is are identical and the sandwich reduces to (2.13). In that case the problem is less complex and involves the estimation of the density, or its reciprochal, at a particular covariate value. However, in the non-iid case the task is much more involved (see Koenker, 2005, Chapter 3).

2.4 Test For Heteroscedasticity

In quantile regression problems it is often useful to consider how the estimated regression parameters $\hat{\boldsymbol{\beta}}_\tau$, for a given design matrix $\boldsymbol{X}$ vary as a function of τ, $\tau \in (0,1)$. In classical linear regression, it is normally assumed that the covariate effects shift the location of the response distribution, but does not change its scale or shape. However, in many quantile regression problems the slope estimates ordinarily changes across the quantiles. Thus, this a basic problem of inference in quantile regression require tests for the regression parameter estimates to be tested for equality across quantiles.

In the linear model with iid errors the conditional quantile function can be written as

$$\mathcal{Q}(\tau|\boldsymbol{x}) = \boldsymbol{x}^{\top}\boldsymbol{\beta} + \mathcal{Q}_\epsilon(\tau) \tag{2.17}$$

where $\boldsymbol{\beta}$ is the parameter vector and $\mathcal{Q}_\epsilon(\tau)$ is the quantile function of the error distribution. In this case,

$$\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + \left(\mathcal{Q}_\epsilon(\tau), 0, \ldots, 0\right)^{\top}$$

influences only the location of the conditional distribution $F(y|x)$ making the quantile functions parallel in the iid case.

When we move from the assumption of iid errors we can consider a more general model of heteroscedasticity,

$$Y = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x})\epsilon \tag{2.18}$$

where $\mu(\boldsymbol{x})$ is the conditional mean of the *regression process*, $\sigma(\boldsymbol{x})$ the conditional scale, and ϵ is an error term independent of x from a distribution with a quantile function $\mathcal{Q}_\epsilon(\tau)$. Considering this, the conditional quantile functions of Y are then

$$\mathcal{Q}_Y(\tau|\boldsymbol{x}) = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x})\mathcal{Q}_\epsilon(\tau), \tag{2.19}$$

which under the assumption that both μ and σ are linear functions of x can be written as

$$\mathcal{Q}_Y(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta} + (\mathbf{1} + \boldsymbol{x}^\top\boldsymbol{\gamma})\mathcal{Q}_\epsilon(\tau) \tag{2.20}$$

for some $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^k \times \mathbb{R}^k$. This *linear scale* model of heteroscedasticity (2.20) is a special case of the models with linear conditional quantile functions. In the case of the iid errors, conditional quantile functions are parallel hyperplanes in k space and the slope coeficients of all regression quantiles coverge in probability to the same vector as $\boldsymbol{\gamma} = \sigma e_1 = \sigma(1, 0, \ldots, 0)^\top$ for $\sigma \geqslant 0$ (Koenker and Bassett, 1982).

Providing that we adopt this linear scale model of heteroscedasticity (2.20), but rather than a fixed vector $\boldsymbol{\gamma}$ which determines the scale parameter, we consider a sequence $\{\gamma_n\}$ that depends on sample size, and if we would consider Theorem 2.3.2 to which we add to the existing two assumptions another one related to the *Scale* problem (Koenker and Bassett, 1982):

1. ***Assumption 1*** *(Density): The error distribution, F, has continuous and strictly positive density, f, for all z such that $0 < F(z) < 1$;*

2. ***Assumption 2*** *(Design): The sequence of design points $\{x_i\}$, satisfies $n^{-1}\sum x_i x_i' \to D$, a positive definite matrix;*

3. ***Assumption 3*** *(Scale): The sequence of scaling functions takes the form $\sigma_n(x) = 1 + x\gamma_n$, where $\gamma_n = \gamma_o/\sqrt{n}$, for some fixed $\gamma_o \in \mathbb{R}^k$;*

$\hat{\boldsymbol{\beta}}(\tau)$ converges in probability to $\boldsymbol{\beta}(\tau)+\mathcal{Q}_\epsilon(\tau)\boldsymbol{\gamma}$ and the slope coefficients depend on τ. By appropriate choice of $\boldsymbol{\gamma} = O(1/\sqrt{n})$, the covariance matrix of $\hat{\boldsymbol{\beta}}(\tau)$ is independent of $\boldsymbol{\gamma}$, thus enabling relevant hypothesis tests of homogenisity to be carried out.

2.4.1 General Linear Hypotheses

Traditional regression analysis is focused on the conditional mean of the response variable y, given x and in general *assumes* that the conditional quantile functions are all parallel to one another. This assumption implies that the covariate effects the location, shift of the response distribution, but do not change its scale parameters, which suggests that the slope coefficients of different quantile regression functions are identical. However, such model assumptions are not always met in the practice in which quantile regression slope estimators often very considerably across different quantile functions. Therefore, an obvious and basic problem of inference in quantile regression involves testing for equality of slope parameters across quantiles.

We can consider the general linear hypotheses on the vector:

$$\hat{\boldsymbol{\zeta}}_n = \left(\hat{\boldsymbol{\beta}}_n\left(\tau_1\right)^\top, ..., \hat{\boldsymbol{\beta}}_n\left(\tau_m\right)^\top\right)^\top$$

of the following form:

$$H_0 \colon \boldsymbol{R\zeta} = r$$

Under the conditions of Theorem 2.3.2 we have the following important result:

Theorem 2.4.1. *(Koenker and Bassett (1982, Thm 4.1)) Under the null hypothesis H_0, the test statistic,*

$$T_n = n(\boldsymbol{R}\hat{\boldsymbol{\zeta}} - r)^\top \left[\boldsymbol{R}(\boldsymbol{\Omega} \otimes \boldsymbol{Q}_0^{-1})\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{R}\hat{\boldsymbol{\zeta}} - r) \tag{2.21}$$

is asymptotically non-central chi-square with rank (R) degrees of freedom and non-centrality parameter

$$\eta = (\boldsymbol{R}(\mathcal{Q}_\epsilon(\tau) \otimes \boldsymbol{\gamma}_0))^\top \left[\boldsymbol{R}(\boldsymbol{\Omega} \otimes \boldsymbol{Q}_0^{-1})\boldsymbol{R}^\top\right]^{-1} (\boldsymbol{R}(\mathcal{Q}_\epsilon(\tau) \otimes \boldsymbol{\gamma}_0)), \tag{2.22}$$

where $\boldsymbol{\Omega}$ is an $m \times m$ matrix composed of the following elements:

$$\omega_{ij} = \frac{(\tau_i \wedge \tau_j - \tau_i\tau_j)}{f(F^{-1}(\tau_i))f(F^{-1}(\tau_j)}.$$

At every quantile the slope parameters are equal in the homoscedastic situation. The parameter vector $\boldsymbol{\beta}$ can be partitioned $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2)^\top$. Similarly partitioning $\boldsymbol{X} = \Big[\mathbf{1}\vdots\boldsymbol{X}_2\Big]$, and setting $r = 0$, gives

$$\boldsymbol{R}_\Delta = \boldsymbol{\Delta} \otimes \boldsymbol{\Phi} \tag{2.23}$$

where $\boldsymbol{\Delta}$ is an $(m-n) \times m$ matrix with typical element $\boldsymbol{\Delta}_{ij} = \delta_{ij} - \delta_{i(j-1)}$, δ_{ij} is the Kronecker delta, and $\boldsymbol{\Phi} = \Big[\boldsymbol{O}\vdots\boldsymbol{I}_{k-1}\Big]$. Thus, for example, when $m = 2$, $\boldsymbol{\Delta} = [1,\ -1]$ and when $m = 3$

$$\boldsymbol{\Delta} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}.$$

Then $\boldsymbol{R}\hat{\boldsymbol{\zeta}} = [\boldsymbol{\Delta} \otimes \boldsymbol{\Phi}]\,\hat{\boldsymbol{\zeta}} = \big(\hat{\beta}_2(\tau_1) - \hat{\beta}_2(\tau_2), \ldots, \hat{\beta}_2(\tau_{m-1}) - \hat{\beta}_2(\tau_m)\big)$,

$$\boldsymbol{R}(\boldsymbol{\Omega} \otimes \boldsymbol{Q}_0^{-1})\boldsymbol{R}^\top = \boldsymbol{\Delta}\boldsymbol{\Omega}\boldsymbol{\Delta}^\top \otimes \boldsymbol{\Phi}\boldsymbol{Q}_0^{-1}\boldsymbol{\Phi}^\top,$$

and

$$\begin{aligned} \eta &= (\boldsymbol{q}_\Delta \otimes \boldsymbol{\Phi}\boldsymbol{\gamma}_0)^\top \left[\boldsymbol{\Delta}\boldsymbol{\Omega}\boldsymbol{\Delta}^\top \otimes \boldsymbol{\Phi}\boldsymbol{Q}_0^{-1}\boldsymbol{\Phi}^\top\right]^{-1} (\boldsymbol{q}_\Delta \otimes \boldsymbol{\Phi}\boldsymbol{\gamma}_0) \\ &= \boldsymbol{q}_\Delta^\top(\boldsymbol{\Delta}\boldsymbol{\Omega}\boldsymbol{\Delta}^\top)^{-1}\boldsymbol{q}_\Delta \cdot (\boldsymbol{\Phi}\boldsymbol{\gamma}_0)^\top(\boldsymbol{\Phi}\boldsymbol{Q}_0^{-1}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}\boldsymbol{\gamma}_0 \end{aligned}$$

where $\boldsymbol{q}_\Delta = \big(\mathcal{Q}_\epsilon(\tau_1) - \mathcal{Q}_\epsilon(\tau_2), \ldots, \mathcal{Q}_\epsilon(\tau_{m-1}) - \mathcal{Q}_\epsilon(\tau_m)\big)$. Thus, in the homoscedastic case, $\boldsymbol{\gamma}_0 = \mathbf{0}$, T is asymptoticaly central χ^2 with $(m-1)(k-1)$ degrees of freedom. The power of the test depends as expected on the design, $\boldsymbol{\gamma}_0$, $\boldsymbol{q}_\Delta$, and, $\boldsymbol{\Omega}^{-1}$ with $n(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, so

$$n\left[\boldsymbol{\Phi}\boldsymbol{Q}_0^{-1}\boldsymbol{\Phi}^\top\right]^{-1} = (\boldsymbol{X}_2^\top\boldsymbol{X}_2 - n\bar{\boldsymbol{x}}_2^\top\bar{\boldsymbol{x}}_2).$$

One of the disadvantage of the above test statistic (2.21) is the required estimation of the nuisace parameters $1/f(F^{-1}(\tau_1))$ and $1/f(F^{-1}(\tau_2))$ as, to be of any practical value for hypothesis testing, $\boldsymbol{\Omega}$ must be replaced with a consistent estimator $\hat{\boldsymbol{\Omega}}$.

2.4.2 Two-Quantiles Problem

Koenker in his book (2005, Chapter 3.3) illustrates the application of the Wald Test as a test of equality between the interquartile ranges of the two samples, as they may

be consider to be a test of the null hypothesis of hommogeneity. Suppose we consider a standard two-sample problem, $y = (y_1, y_2)^\top$, with the following design

$$X = \begin{bmatrix} 1_{n_1} & \mathbf{0} \\ \mathbf{0} & 1_{n_2} \end{bmatrix},$$

in which $x_i = 0$ in the first sample with n_1 number of observations and $x_i = 1$ in the second sample with n_2 observations. The τ^{th} regression quantile estimate, $\hat{\boldsymbol{\beta}}(\tau) = \big(\hat{\beta}_1(\tau),\ \hat{\beta}_2(\tau)\big)$, has the property that $\hat{\beta}_1(\tau)$ is a τ^{th} sample quantile from sample 1 and that $\hat{\beta}_2(\tau)$ is a τ^{th} sample quantile from sample 2, both consisting of the first n_1 and the last n_2 observations respectively. Using the standard notation for the empirical quantile function, $\hat{\mathcal{Q}}_i(\tau)$, for the two samples we have

$$\hat{\boldsymbol{\beta}}(\tau) = (\hat{\mathcal{Q}}_1(\tau),\ \hat{\mathcal{Q}}_2(\tau)).$$

Following Koenker and Bassett (1982) this standard two-sample problem can be written more conveniently by forming the transformation $\boldsymbol{XA}$, as a simple regression model which has a design matrix with an intercept term and the single $(0, 1)$ binary explanatory variable;

$$\tilde{\boldsymbol{X}} = \boldsymbol{XA} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \vdots & \vdots \\ \mathbf{1} & \mathbf{1} \end{bmatrix}.$$

The results established in Koenker and Bassat (1982) give rise to the following estimates:

$$\tilde{\boldsymbol{\beta}}(\tau) = \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}(\tau) = \begin{pmatrix} \hat{\mathcal{Q}}_1(\tau) \\ \hat{\mathcal{Q}}_2(\tau) \quad - \quad \hat{\mathcal{Q}}_1(\tau) \end{pmatrix}.$$

Consequently examining the issue of heteroscedasticity reduces to an hypothesis test of the form

$$\begin{aligned} \beta_2(\tau_2) - \beta_2(\tau_1) &= \Big(\mathcal{Q}_2(\tau_2) - \mathcal{Q}_1(\tau_2)\Big) - \Big(\mathcal{Q}_2(\tau_1) - \mathcal{Q}_1(\tau_1)\Big) \\ &= \Big(\mathcal{Q}_2(\tau_2) - \mathcal{Q}_2(\tau_1)\Big) - \Big(\mathcal{Q}_1(\tau_2) - \mathcal{Q}_1(\tau_1)\Big) \\ &= 0. \end{aligned}$$

In this case, the τ^{th} quantile regression estimate of the *slope* parameter β_1 is the difference between the τ^{th} sample quantiles of the two samples.

$$y_i = \beta_0(\tau_j) + \beta_1(\tau_j)x_i + e_i, \qquad j = 1, 2$$

Thereby, a test of the equality of the *slope* parameters across quantiles τ_1 and τ_2 is a test of the hypothesis that the $(\tau_2 - \tau_1)$ interquantile ranges are equal for the two samples:

$$\beta_1(\tau_2) - \beta_1(\tau_1) = 0.$$

According to (2.13) the asymptotic variance of $\hat{\beta}_1(\tau_2) - \hat{\beta}_1(\tau_1)$ is given by

$$\sigma^2(\tau_1, \tau_2) = \left[\frac{\tau_1(1-\tau_1)}{f^2(\xi_{\tau_1})} - 2\frac{\tau_1(1-\tau_2)}{f(\xi_{\tau_1})f(\xi_{\tau_2})} + \frac{\tau_2(1-\tau_2)}{f^2(\xi_{\tau_2})}\right]\left[\frac{n_1+n_2}{(n_1+n_2)n_2 - n_2^2}\right],$$

where $\xi_{\tau_i} = F^{-1}(\tau_i)$, n_1 and n_2 are the sizes of the two samples and the test of the null hypothesis can be based on the statistic

$$T_n = \frac{\hat{\beta}_1(\tau_2) - \hat{\beta}_1(\tau_1)}{\hat{\sigma}(\tau_1, \tau_2)}, \tag{2.24}$$

which is $\sim \mathcal{N}(0, 1)$ asymptotically.

As mentioned earlier, the statistics T_n given in (2.21) is asymptotically χ^2_q under H_0, where q is the number of linearly independent restrictions. A practical disadvantage of this test statistic is the required estimation of the nuisance parameters $1/f(F^{-1}(\tau_1))$ and $1/f(F^{-1}(\tau_2))$. However, the flexibility of the formulation of the hypothesis accommodates a wide choice of testing situations, from a simple test on a single quantile to joint tests of regression coefficients involving several distinct quantiles (Koenker, 2005, Chapter 3). A detailed application of the Wald Test can be seen in the Appendix A.

2.5 Asymptotics for Inference

When inference on a particular quantile is required, we have to chose the method to use. In general, the methods are classified based on direct estimation of the asymptotic covariance matrix and those based on some form of the bootstrap. Methods

based on the estimation of the covariance matrix are not straight forward. They depend on the estimation of the reciprocal of the error density at the quantile of interest, a quantity known as the *sparsity function*. This quantity reflects the density of observations near the given quantile and it is understandable that the precision of quantile estimates, $\hat{\boldsymbol{\beta}}_n(\tau)$, should depend on it. If the data are very sparse at the given quantile it will be difficult to estimate accurately. Nonetheless, in the opposite situation, when there are many observations near the quantile of interest and therefore the sparsity is low, then the quantile will be estimated more acurately. It is clear to see that the precision of $\hat{\boldsymbol{\beta}}_n(\tau)$ depends entirely on the sample information in the neighborhood of the τ^{th} quantile. In order to estimate the accuracy of the τ^{th} quantile regression estimate, we need to estimate the sparsity at the given quantile of interest

$$s(\tau) = \left[f\left(F^{-1}(\tau)\right)\right]^{-1} \tag{2.25}$$

which takes us into the field of the density estimation and smoothing. Subsequently, the bootstrapping methodology can be seen as a way forward in that such that the sparsity no longer has to be estimated, thus eliminating one nuisance problem (Koenker, 2005, Chapter 3).

2.5.1 Sparsity Estimation

Estimation of the reciprocal of a density function has been studied before Koenker and Bassett's (1978) introduction to *quantile regression* and before the relevance of this quantity to the asymptotic accuracy of a quantile of interest has been recognised. It was named the *"sparsity function"* by Tukey (1965). For quantile models with iid errors, this parameter can be seen as equivalent to the standard deviation of the error in the least-squres estimation problem in the simple regression model, with iid errors.

Siddiqui's (1960) idea on how the sparsity function, $s(\tau)$, could be estimated has received the most attention in the literature and is often regarded as simplest (Koenker, 1994). It could be observed as an inverted image of the density estimate. By differentiating the identity $F\left(F^{-1}(t)\right) = t$, the sparsity function can be derived

from the quantile function as follows:

$$\frac{d}{dt}F^{-1}(t) = s(t). \tag{2.26}$$

Using the suggestion from Siddiqui (1960) for constructing confidence intervals for a univariate sample quantile, $s(t)$ can be estimated. Having the estimate $\hat{F}_n^{-1}(\cdot)$ of F^{-1} and a bandwidth h_n which tends to zero as $n \to \infty$ we can estimate the sparsity as follows:

$$\hat{s}_n(t) = \left[\hat{F}_n^{-1}(t+h_n) - \hat{F}_n^{-1}(t-h_n)\right]/2h_n. \tag{2.27}$$

This Siddiqui's solution, however, poses some further problems of which the first and foremost is the choice of the bandwidth.

One of the possible bandwidth choices has been suggested by Bofinger (1975). Based on minimization of the mean square error of the density estimator, Bofinger shows that the h_B bandwidth can be expressed as follows:

$$h_B = n^{-1/5}\left[\frac{4.5s^2(t)}{(s''(t))^2}\right]^{1/5}. \tag{2.28}$$

Another option would be to choose the h_{HS} bandwidth suggested by Hall and Sheather (1988) based on Edgeworth expansions for studentised univariate sample quantiles. They suggest

$$h_{HS} = n^{-1/3}z_\alpha^{2/3}\left[\frac{1.5s(t)}{s''(t)}\right]^{1/3} \tag{2.29}$$

where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha/2$ for the construction of $1-\alpha$ confidence intervals (Koenker, 2005, Chapter 4.10).

Figure 2.5 illustrates the bandwidth parameter h_n depending on the sample size calculated using Bofinger and Hall and Sheather's bandwidth respectively for three distinct quantiles: $\tau = 0.25,\ 0.5$ and 0.95. Observing Figure 2.6 we can see that the h_n sequences for τ and $(1-\tau)$ for symmetric F, such as Normal distribution, are evidently identical. We can notice that h_n has a bigger value when using Hall and Sheather's bandwidth for the small sample sizes and as the sample size increases to moderate and large Hall and Sheather's bandwidth produces smaller h_n than Bofinger's bandwidth.

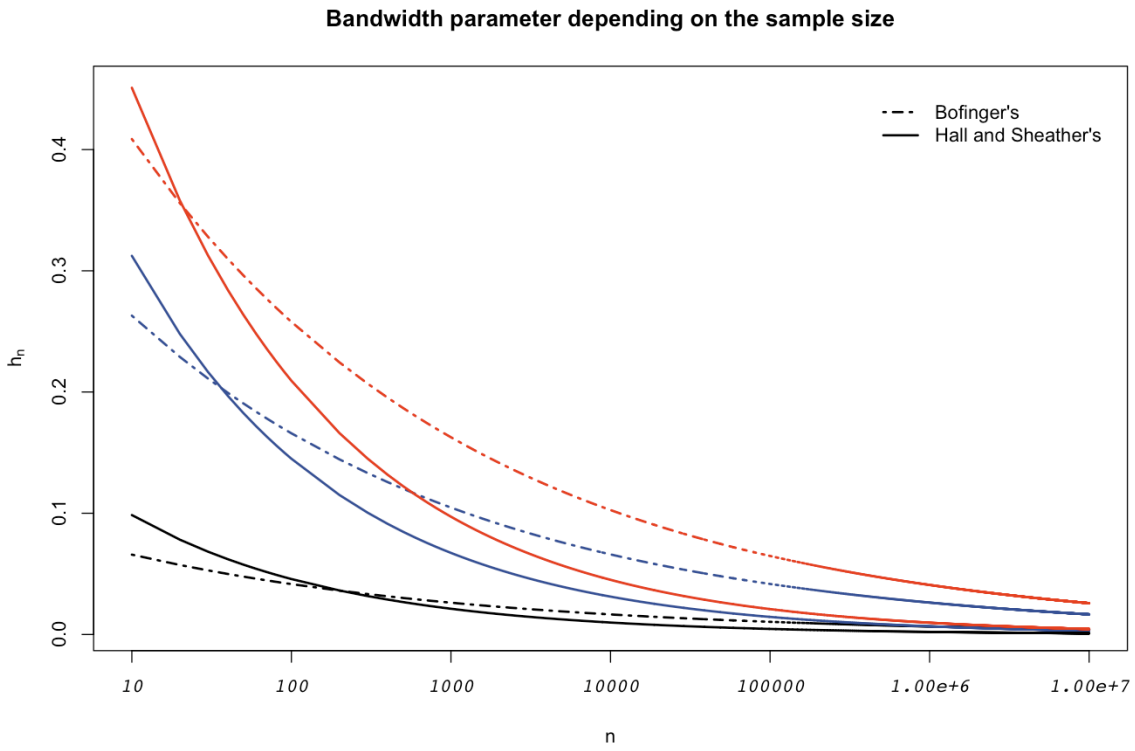


Figure 2.5: *Bandwidth parameter h_n. Dashed line present Bofinger's bandwidths and solid lines Hall and Sheathers's bandwidths for different quantiles of interest (blue, red and black for $\tau = 0.25$, $\tau = 0.5$ and $\tau = 0.95$ respectively). Hall and Sheathers's bandwidths are calculated for $\alpha = 0.05$.*

However, we may find all of a sudden ourself in a loop as we suppose to estimate sparsity $s(t)$ by using a chosen bandwidth for which we need the unknown $s(t)/s''(t)$ ratio. Of course, if we would know $s(t)$ and $s''(t)$ we would not need neither h_B nor h_{HS}. In the absence of $s(\cdot)$ we can calculate a chosen bandwidth for some typical distributional shape, like Gaussian, as fortunately $s(t)/s''(t)$ is not very sensitive to F (Koenker, 1994).

In general,

$$\frac{s(t)}{s''(t)} = \frac{f^2}{2(f'/f)^2 + [(f'/f) - f''/f]}$$

and if we would plug in the Gaussian density,

$$\frac{f'}{f} F^{-1}(t) = \Phi^{-1}(t)$$

the optimal Bofinger's bandwidth would become

$$h_B = n^{-1/5} \left[\frac{4.5\phi^4(\Phi^{-1}(t))}{(2\Phi^{-1}(t)^2 + 1)^2} \right]^{1/5},$$

and for Hall and Sheather's bandwidth would yield to

$$h_{HS} = n^{-1/3} z_\alpha^{2/3} \left[\frac{1.5\phi^2(\Phi^{-1}(t))}{(2(\Phi^{-1}(t))^2 + 1)} \right]^{1/3}.$$

Figure 2.6 illustrates different bandwidth choice rules for different τ's for a sample size of $n = 500$, based on a Gaussian distribution.

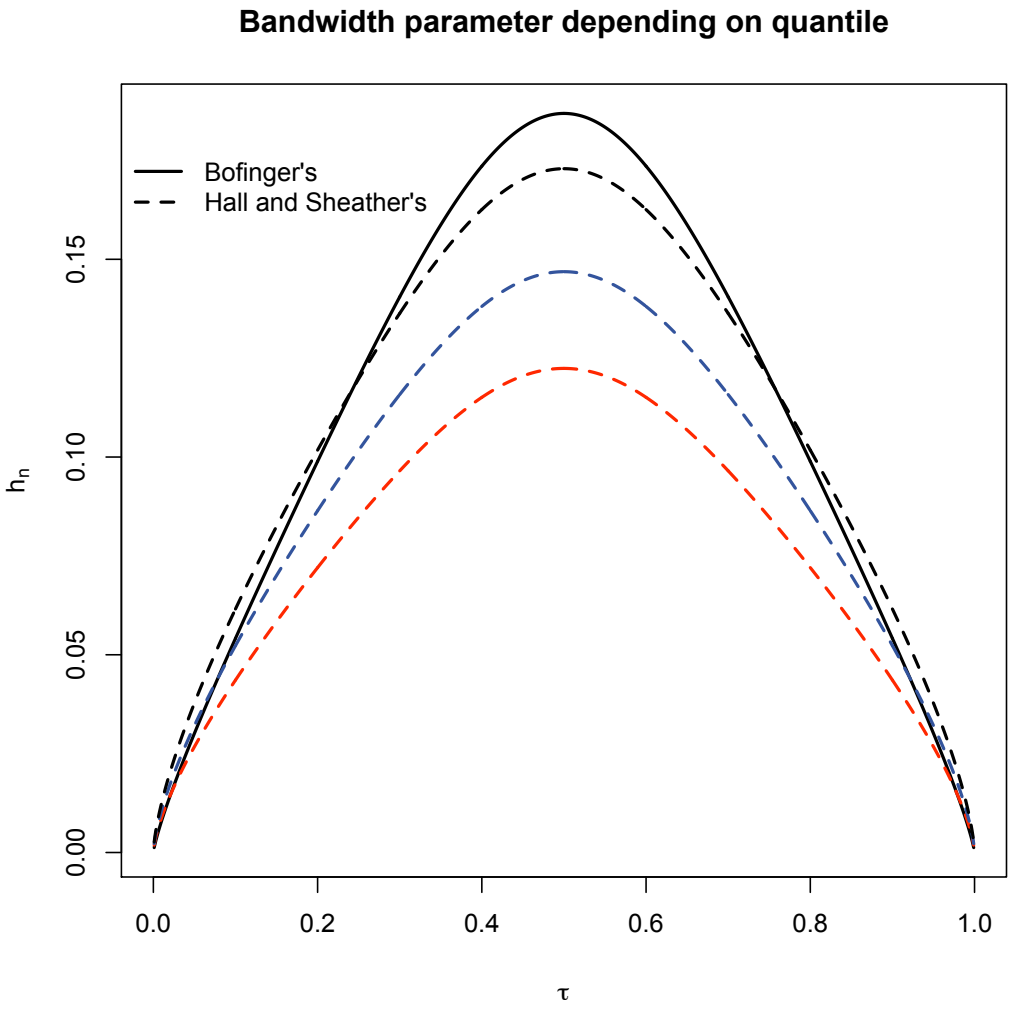


Figure 2.6: *Bandwidth parameter h_n. Solid line present Bofinger's bandwidth and dashed lines Hall and Sheathers's bandwidth for different levels of α (red, blue and black for $\alpha = 0.05$, $\alpha = 0.01$ and $\alpha = 0.001$ respectively) of a sample size $n = 500$.*

Note that estimation of the sparsity $s(t)$ is far more simpler 'problem' for iid quantile regression models, as $s(t)/s''(t)$ constant is location-scale invariant and is affected only by the *shape* of the distribution.

Having looked at the two possible ways of determining the bandwidth parameter h_n, the question that still remains is how to calculate the estimated empirical quantile function $\hat{\mathcal{Q}}(\tau) = \hat{F}^{-1}(\tau)$ in (2.27). As suggested by Bassett and Koenker (1982), one approach would be to make use of the empirical quantile function. This approach uses

$$F_n^{-1}(\tau) = \bar{\boldsymbol{x}}^\top \hat{\boldsymbol{\beta}}_n(\tau), \tag{2.30}$$

where $\bar{x} = n^{-1} \sum x_i$ and $\hat{\boldsymbol{\beta}}(\cdot)$ is the usual regression quantile process. The functions

$$\hat{\mathcal{Q}}_y(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}(\tau)$$

represent a family of conditional quantile functions for the response variable y. We can estimate the conditional quantile function $\hat{\mathcal{Q}}_y(\tau|\boldsymbol{x})$ of y at any fixed x, but as mentioned earlier, the precision of the quantile depends upon the sparsity at the given x at which we evaluate the expression. Therefore, in the cases of symmetric, unimodal *pdf*, it is reasonable to expect that the precision is maximised at $x = \bar{x}$. In order to employ this approach when estimating the *sparsity* $\hat{s}(\tau)$ we need to be sure that $\hat{F}^{-1}$ satisfies the essential monotonicity requirement of a quantile function (Koenker, 1994).

Theorem 2.5.1. *(Bassett and Koenker (1982, Thm 2.1)), The sample paths of* $\hat{\mathcal{Q}}(\tau|\bar{\boldsymbol{x}})$ *are nondecreasing, left-continuous, jump function on* $(0,\ 1)$.

In the location models the jumps in the empirical quantile function occur at equally spaced points $\{i/n :\ i = 1, \ldots, n\}$ on $(0,\ 1)$, where by the location model we consider for some vector of independent random variables $\boldsymbol{Y} = (Y_1, Y_2, \ldots,\ Y_n)$ the following model

$$Pr[Y_i \leqslant y|\boldsymbol{x}_i] = F\left(y - \sum_{j=1}^{p} \boldsymbol{x}_{ij}\boldsymbol{\beta}_j\right), \quad i = 1, \ldots, n,$$

in which $\boldsymbol{\beta}$ is a scalar and $\mathbf{X} = \mathbf{1}_n$. If we would now examine a linear model in which $\mathbf{X}$ is a design matrice of size $p \times n$, the jumps in $\hat{\mathcal{Q}}(\cdot)$ occur at random points on $(0,\ 1)$ (see Figures 2.7 and 2.8) depending upon the design as well as the realisation of $\boldsymbol{Y}$ (Bassett and Koenker, 1982).

Let us consider a simple bivariate sample of only seven observations consisting of (x, y) pairs $(1, 2)$, $(2, 1)$, $(3, 5)$, $(5, 2)$, $(6, 8)$, $(8, 6)$ and $(9, 9)$. Figure 2.7 illustrates the regression quantiles for $\tau \in (0.05, 0.25, 0.5, 0.75, 0.95)$ and their unique solutions to (2.8) at $\hat{\boldsymbol{\beta}}$ are given in Table 2.1.

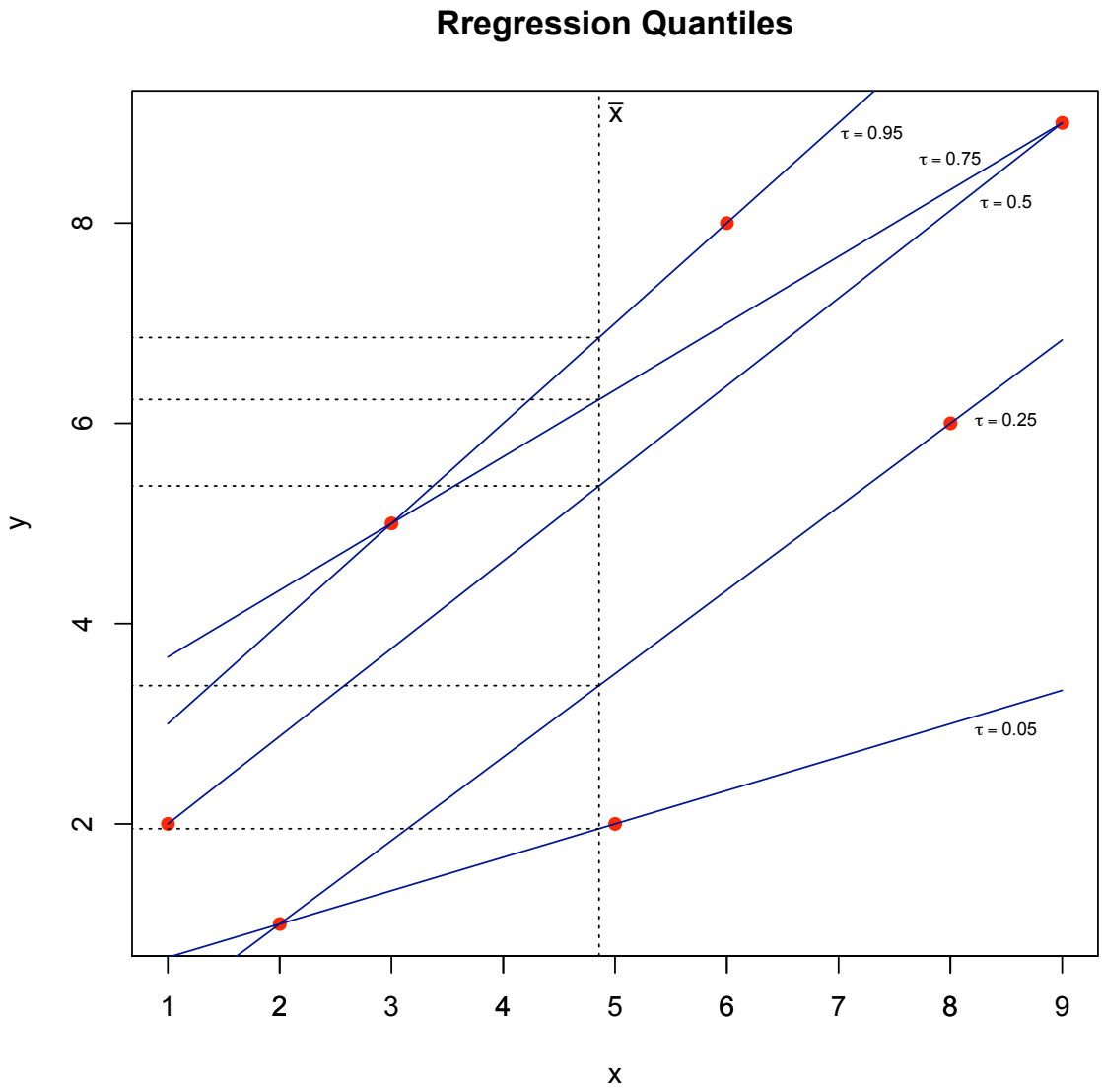


Figure 2.7: *Regression Quantiles for Bivariate Example With Seven Observations. The blue, solid lines are quantile regression estimates for* $\tau \in (0.05, 0.25, 0.5, 0.75, 0.95)$, *intersected by the vertical dashed line located at* $\bar{x} = 4.857$. *The ordinates at the intersections determine the empirical quantile function for this data set, illustrated in Figure 2.8.*

Thus, according to Theorem 2.3.2, in the linear models with iid error the asymptotic behavior of linear combinations of regression quantiles is similar to the large

$\hat{\boldsymbol{\beta}}$	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
$\hat{\beta}_0$	0.333	-0.667	1.125	3.000	2.000
$\hat{\beta}_1$	0.333	0.833	0.875	0.667	1.000

Table 2.1: Quantile regression estimates for the simple bivariate example

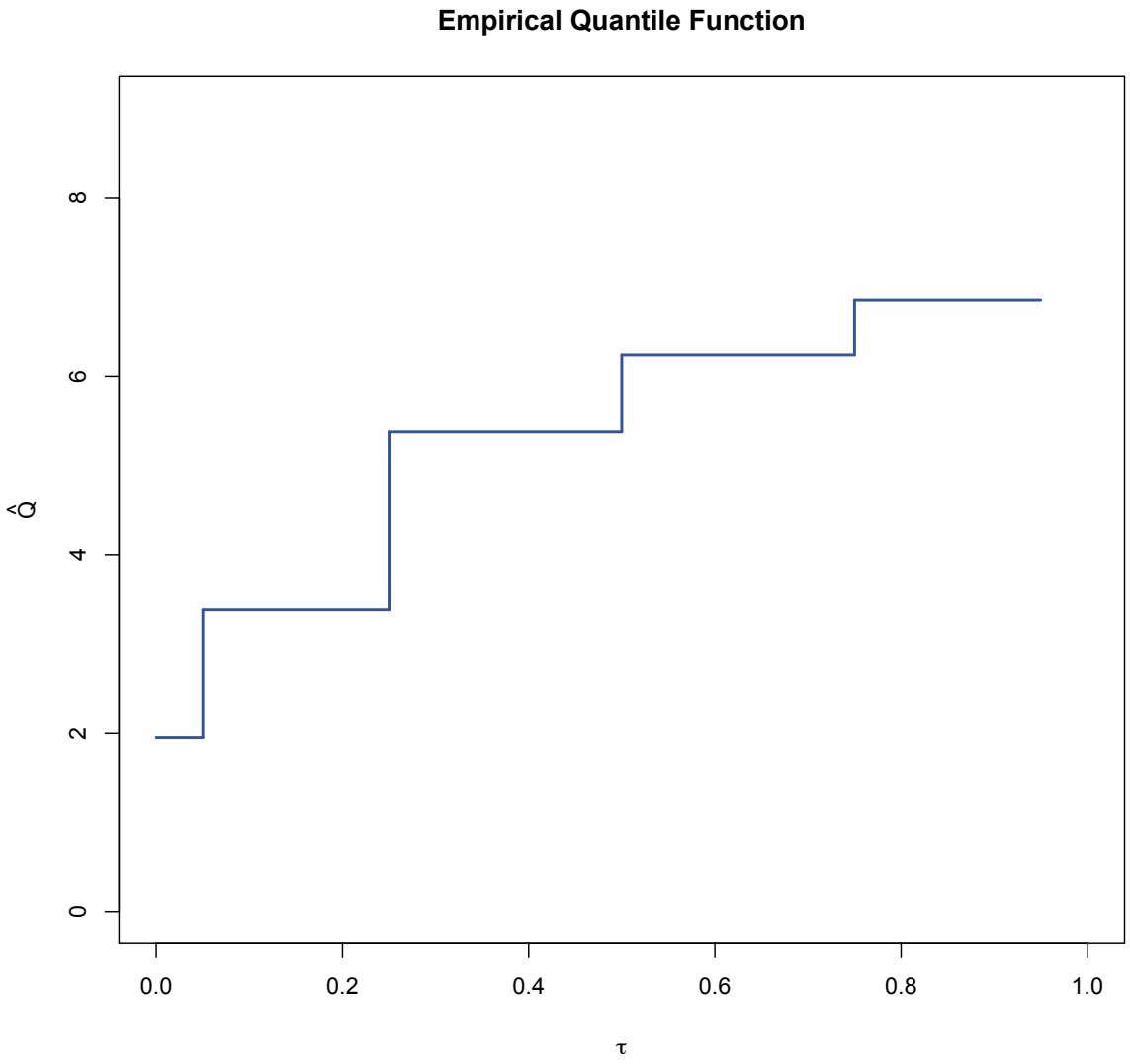


Figure 2.8: *The Empirical Quantile Function. The line represent* $\hat{\mathcal{Q}}(\tau|\bar{x})$ *based on the data in Figure 2.7.*

sample theory of ordinary sample quantiles in the one sample, location model.

Theorem 2.5.2. *(Bassett and Koenker (1982, Thm 3.2)) The finite dimensional*

distributions of the random function

$$Z_n(\tau|\boldsymbol{x}) = \sqrt{n}(\hat{\mathcal{Q}}(\tau|\boldsymbol{x}) - \mathcal{Q}(\tau|\boldsymbol{x}))$$

are asymptotically Gaussian with zero mean and covariance matrix $\boldsymbol{x}\boldsymbol{Q}_{\mathbf{0}}^{-\mathbf{1}}\boldsymbol{x}^{\top}\boldsymbol{\Omega}$. *Specialising to* $\boldsymbol{x} = \bar{\boldsymbol{x}}$ *we have* $Z_n(\tau|\bar{\boldsymbol{x}}) \to G(0,\ \boldsymbol{\Omega})$.

Proof. This is immediate from the definition of $\hat{\mathcal{Q}}$, Theorem 2.3.2, and the identity $\bar{\boldsymbol{x}}\boldsymbol{Q}_{\mathbf{0}}^{-\mathbf{1}}\bar{\boldsymbol{x}}^{\top} = 1$ which follows from the fact that $\mathbf{1}$ is in the column space of $\mathbf{X}$. □

As a result of this, it is possible to obtain similar convergence results for the empirical quantile function $\hat{\mathcal{Q}}(\tau)$, and for the corresponding estimator of the error distribution $\hat{F}(x) = \hat{\mathcal{Q}}^{-1}(x)$ (Portnoy, 2010). From Theorem 2.5.1 we have that $\hat{F}(y) \equiv \hat{F}(y|\bar{\boldsymbol{x}})$ is a non-decreasing, right-continuous, jump function on $\mathbf{R}^1$. Versions of $\hat{\mathcal{Q}}(\cdot)$ and $\hat{F}(\cdot)$ with continuous sample path could be easily constructed (Bassett and Koenker, 1982). Thus, in iid error model where only a scalar estimate of the sparsity function at each τ is required, according to (2.27) we have

$$\hat{s}(\tau) = \left[\hat{\mathcal{Q}}_y(\tau + h_n|\bar{\boldsymbol{x}}) - \hat{\mathcal{Q}}_y(\tau - h_n|\bar{\boldsymbol{x}})\right]/2h_n. \tag{2.31}$$

Figure 2.9 illustrate the function $\hat{\mathcal{Q}}_y(\tau|\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}}^{\top}\hat{\boldsymbol{\beta}}(\tau)$ of conditional quantiles of girls weight at age $1\frac{1}{2}$ years. The dotted lines forming triangle illustrate the estimates of the sparsity function at $\tau = 0.25$ using Bofinger's and Hall-Sheather's bandwidths of 0.073 and 0.037 providing the sparsity estimates $\hat{s}(\tau = 0.25) = 6.377$ and $\hat{s}(\tau = 0.25) = 6.178$ respectively. Alternative, more simple approach to this for computing F^{-1} would be to use the residual from the quantile regression fit:

$$u_i = y_i - \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}(\tau) \quad i = 1, \ldots, n. \tag{2.32}$$

Let $u_i \colon i = 1, \ldots, n$ be those residuals and $u_{(i)} : i = 1, \ldots, n$ be the corresponding order statistics. We have got that the empirical quantile function based on this "sample" is:

$$\hat{F}^{-1}(\tau) = u(j) \ \text{ for } \ \tau \in \left[\frac{j-1}{n}, \frac{j}{n}\right) \tag{2.33}$$

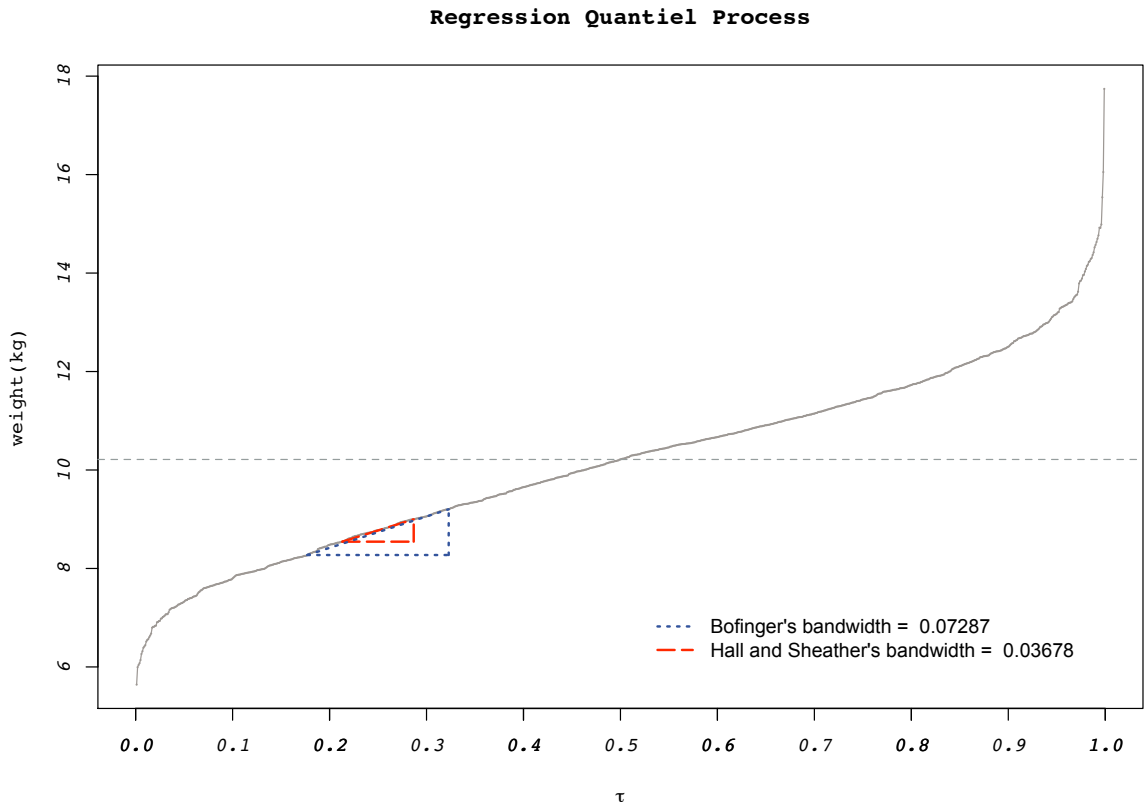


Figure 2.9: *Sparsity Estimation for the Girls Weight, age birth to 36 months. This plot shows* $\hat{\mathcal{Q}}_Y(\tau|\bar{x}) = \bar{x}'\hat{\beta}(\tau)$ *for the girls logarithmic weight. Blue and red dotted triangles depict estimators of the sparsity at the first quartile* $\tau = \frac{1}{4}$ *using Siddiqui method. The estimates of the sparsity are given by the slopes of the hypothenuse of the triangles.*

One way of interpolating this in order to get a piecewise linear version would be to use the following:

$$\tilde{F}^{-1}(\tau) = \begin{cases} u_{(1)} & \text{if } \tau \in \left[0, \frac{1}{2n}\right) \\ \lambda u_{(j+1)} + (1+\lambda)u_{(j)} & \text{if } \tau \in \left[\frac{2j-1}{2n}, \frac{2j+1}{2n}\right),\ j = 1, \ldots, n-1 \\ u_{(n)} & \text{if } \tau \in \left[\frac{2n-1}{2n}, 1\right] \end{cases}$$

where $\lambda = \tau n - j + 1/2$ (Koenker, 1994).

A possible problem that we may encounter when using this residual based approach is when the number of parameters estimated (p) is large relative to n. In that case, we need to ensure that the bandwidth is large enough to avoid the p zero residuals that we would get as a consequence of the quantile fitting procedure. One possible solution to this problem can be drawn from the analogy to the usual

degree-of-freedom correction in estimating σ^2 in least-squares regression (Koenker, 2005, Chapter 4.10). The simplest thing would be to ignore zero residuals from the "sample" of u_is and to regard the effective sample size as $n - p$.

There are many other possible approaches to the estimation of the sparsity parameter and Koenker (2005) in his book (pg.79) provides a short references list of the most comprehensive methods that could be used. For example, we can try to estimate the sparsity by fitting a local polynomial to $\hat{F}^{-1}(t)$ in the neighborhood of τ and use the slope of this fitted function at τ as an estimate. On the other hand, Welsh (1988) suggests a kernel estimation of the sparsity function. His approach could be interpreted as a weighted average of Siddiqui estimates whereby narrow bandwidth are given a greater degree of importance. In Koenker and Bassett (1982), they estimate sparsity function by twice differentiating a smoothed version of $\hat{R}(\tau)$, given by equation (2.4), which denotes the minimum value achieved by the objective function at each regression quantile.

To demonstrate the application of this approach we will use girls weight data set, age birth to 3 years to apply test statistic (2.21) to assess for heteroscedasticity. Observing Figure 2.10 appear the data to be heteroscedastic, as the heavier the babies are at birth, the higher the rate of weight gain is. Table 2.2 presents the results from the seven estimates of the weight-age linear relationship, from which is also clear that as the quantiles increase, so do the gain in weights with the age. Standard errors for constructing 95% confidence intervals of the estimated linear conditional quantile process for these data, presented in Figure 2.11, are calculated using a kernel estimate of the sandwich as proposed by Powell (1991) (see the following section). They are clearly illustrating increase in the slope parameter as value of τ increases. The next question is whether the rate in weight gain with age differs significantly across quantiles. To confirm the visual impression, that the difference is visibly significant, we can focus on the inner two lines, which present the fit for the first ($\tau = 0.25$) and third ($\tau = 0.75$) quartiles and test the the following hypothesis: $H_0 \colon \ \beta_1(0.25) = \beta_1(0.75)$ against $H_1 \colon \ \beta_1(0.25) \neq \beta_1(0.75)$

Figure (2.9) illustrates the function $\hat{\mathcal{Q}}_Y(\tau|\bar{\boldsymbol{x}})$ for this data set, where $\bar{x} = 1.5$.

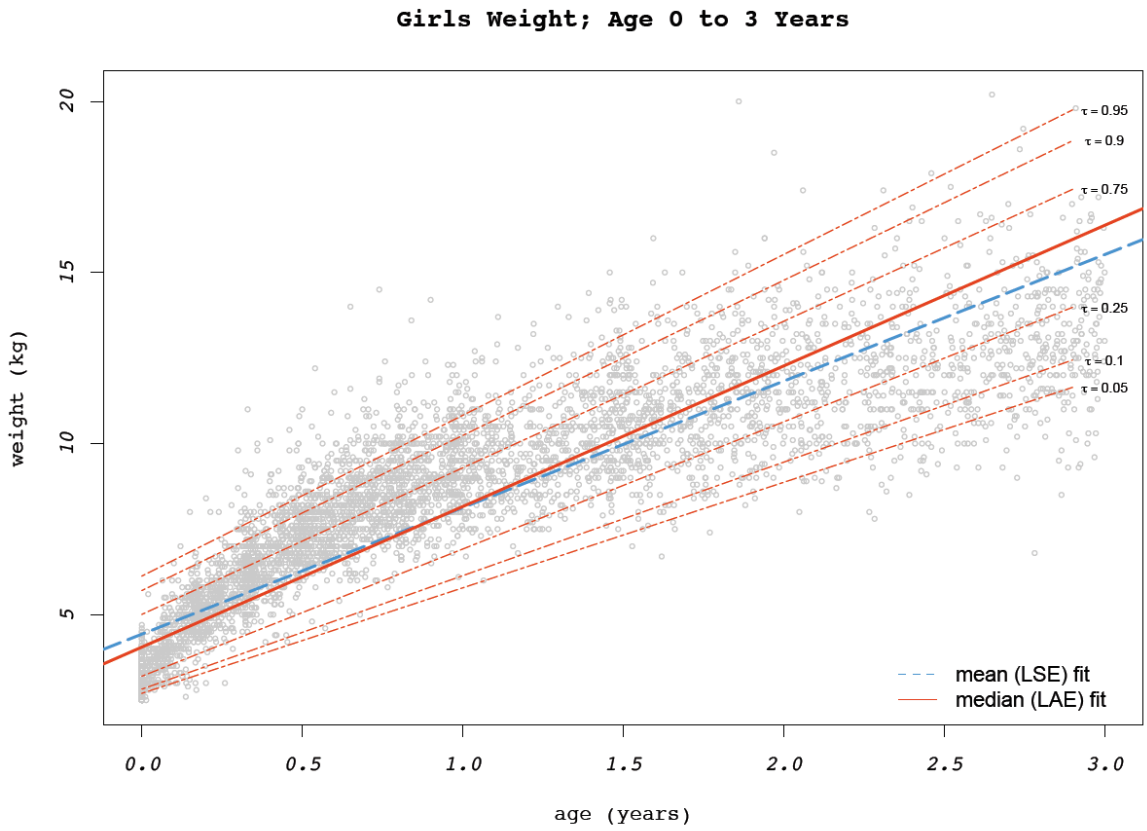


Figure 2.10: *Scatterplot and Quantile Regression Fit for Girls Weight, Age Birth to 36 Months: The plot shows a scatterplot of the girls weight, age birth to 3 years, for a sample of* $6,123$ *observations. Superimposed on the plot are the* $\{0.05,\ 0.01,\ 0.25,\ 0.50,\ 0.75,\ 0.90,\ 0.95\}$ *quantile regression lines in dashed red, the median fit in a solid red line, and the least square estimate of the conditional mean function as the solid blue line.*

$\hat{\boldsymbol{\beta}}/se$	OLS	$\tau = 0.05$	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	$\tau = 0.95$
$\hat{\beta}_0$	4.430	2.700	2.822	3.200	4.051	5.007	5.705	6.126
se	0.028	0.006	0.018	0.026	0.041	0.030	0.034	0.061
$\hat{\beta}_1$	3.698	3.084	3.317	3.718	4.109	4.284	4.533	4.699
se	0.024	0.030	0.032	0.041	0.041	0.038	0.047	0.058
	$R^2 = 0.795$							

Table 2.2: OLS and quantile regression estimates for the following model: $Weight = \beta_0(\tau) + \beta_1(\tau)Age$.

The Hall-Sheather bandwidth for both estimates is 0.037, yielding sparsity estimates of $\hat{s}(0.25) = 6.178$ and $\hat{s}(0.75) = 5.649$. We find the the difference in the slopes is

$$\hat{\beta}_1(0.75) - \hat{\beta}_1(0.25) = 4.284 - 3.718 = 0.566$$

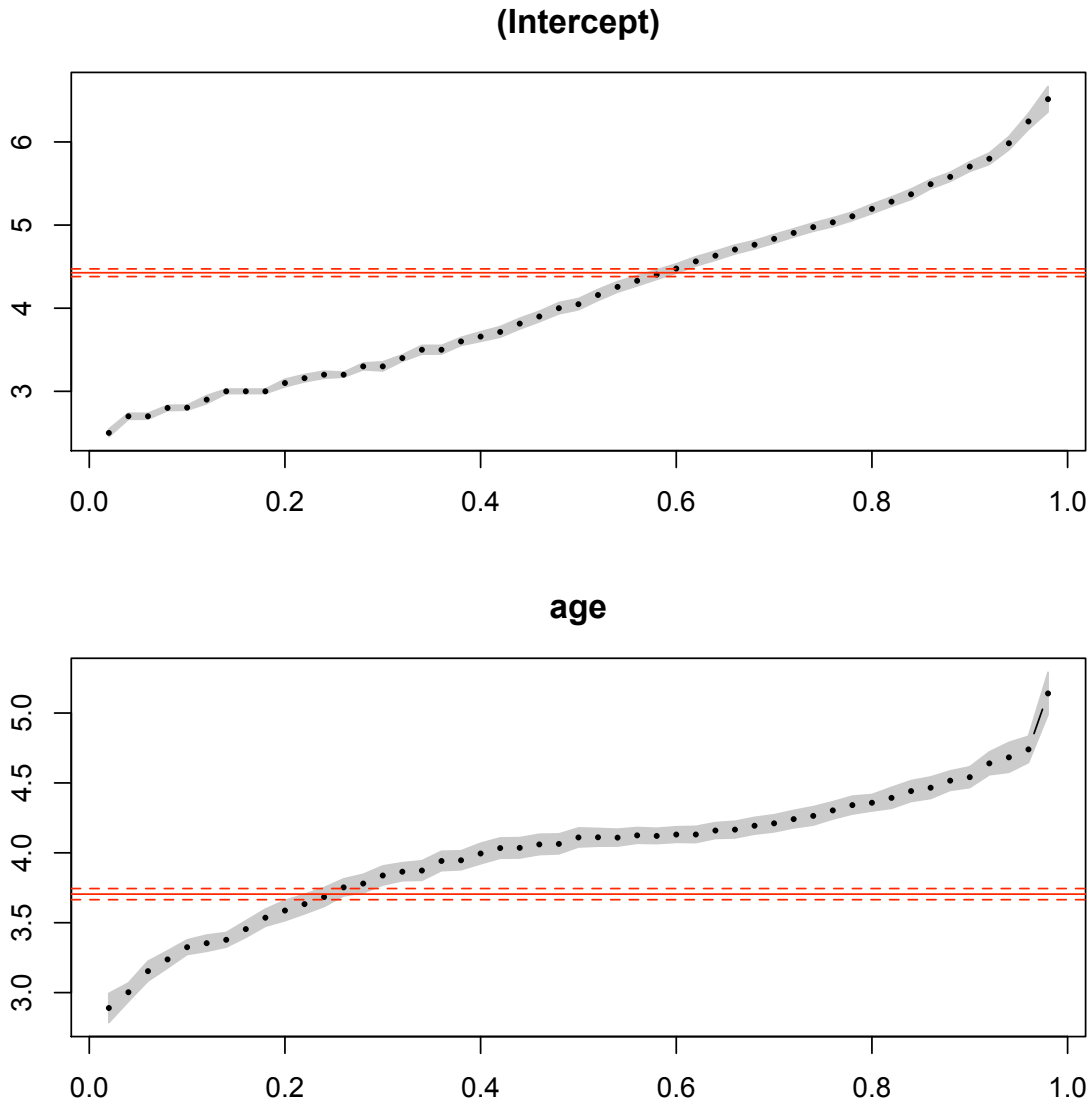


Figure 2.11: *Girls Weight, Age Birth to 3 Years Plots: the intercept and slope of the estimated linear quantile regression lines for the girls weight, age birth to 3 years are ploted as a function of τ. A 95% confidence interval for the quantile regression parameters is indicated by the shaded area. The solid lines indicate the least-square estimates and 95% confidence intervals are represented by the dashed lines in each of the two panels.*

This data set consists of 6, 123 observations, the lower diagonal element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = 0.0002$, and so the test statistic for the equality of the two slopes is 12.428, which has a p-value less than 0.0001 for one-tailed test of the hypothesis of equality of the slopes (see Appendix A).

2.5.2 Estimating The Covariance Matrix: Non-iid Models

Koenker (2005) in his book (pg. 79) provides a review of two methods for estimating the matrix $\boldsymbol{H}_n(\tau)$. In the first, *Hendricks-Koenker Sandwich* approach, assumes linearity in the τ^{th} conditional quantile function. For a chosen bandwidth $h_n \to 0$, mention earlier, the parameters of the $\tau \pm h_n$ quantile function and the density $f_i(\xi_i)$ can be estimated by using the following:

$$\hat{F}_i\left(\xi(\tau)\right) = 2h_n/\boldsymbol{x}_i^\top \left(\hat{\boldsymbol{\beta}}(\tau + h_n) - \hat{\boldsymbol{\beta}}(\tau - h_n)\right).$$

Using this estimate in the expression for H_n would enable the estimation of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}(\tau)$ in the non-iid error model. However, since the quantity

$$d_i = \boldsymbol{x}_i^\top \left(\hat{\boldsymbol{\beta}}(\tau + h_n) - \hat{\boldsymbol{\beta}}(\tau - h_n)\right)$$

is positive at $x = \bar{x} = n^{-1}\Sigma x_i$ *only* and therefore there is no guarantee of positivity for every observation in the sample, which may cause a possible problem with this estimation of $\hat{F}_i(\xi(\tau))$. Koenker (2005, Chapter 3) points out, that since in practice the problem due to "crossing" is infrequent and it happens in the most extreme regions of the design space, $\hat{F}_i$ can be replace by its positive part. Thus, in the rare cases in which $d_i = 0$, where the i^{th} observation is laying on both $\tau \pm h_n$ quantile function lines

$$\hat{F}_i^+ = \max\{0, 2h_n/(d_i - \xi)\},$$

where $\xi > 0$ is a small tolerance parameter.

Another approach to estimating $H_n(\tau)$, that has been implemented in the `quantreg` package in `R`, is *The Powell Sandwich*. To obtain the Powell kernel version of the asymptotic covariance matrix of the quantile regression estimator we would use the following:

$$\hat{\boldsymbol{H}}_n(\tau) = (nh_n)^{-1} \sum K\left(u_i(\tau)/h_n\right)/\boldsymbol{x}_i\boldsymbol{x}_i^\top,$$

where $u_i(\tau) = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$. The kernel function K is a second-order function $K(x) : [0,1] \to \mathbb{R}$ satisfying $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx = \sigma_K^2 \neq 0$ and

h_n is an appropriate bandwidth parameter requiring $h_n \to 0$ and $\sqrt{n}h_n \to \infty$ (for more details see Section 3.3).

Of course, this approach poses a set of questions regarding the choice of the kernel function K and the bandwidth parameter h_n. Powell (1991) suggests:

$$\hat{\boldsymbol{H}}_n(\tau) = \frac{1}{2nc_n} \sum I(|\hat{u}_i| < c_n)\boldsymbol{x}_i\boldsymbol{x}_i^\top.$$

In R Koenker as a default bandwidth implements

$$c_n = \mathcal{K}\left(\Phi^{-1}(\tau + h_n) - \Phi(\tau - h_N)\right),$$

(Koenker, 2005, Chapter 3).

2.6 Rank-Score Process

Rank and order statistics are used to form non-parametric statistical tests. Gutenbrunner and Jurečková (1992) have made an important connection to the classical theory of rank tests by introducing *regression rank-scores* as the counterpart of the univariate ranking idea.

Sorting and ranking the sample observations are the only computations required for the simple linear rank tests and these ideas can be extended to more general models that are linked to the optimization framework suggested by Koenker and Bassett (1978). It is well known that finding the τ^{th} quantile can be expressed as a linear programing problem. The quantile regression setting can be solved as a primal problem given by

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p; u,v\in\mathbb{R}^n_+} \{\tau \mathbf{1}_n^\top u + (1-\tau)\mathbf{1}_n^\top v | \boldsymbol{X}\boldsymbol{\beta} + u - v = y\}. \tag{2.34}$$

The dual of this quantile regression linear programming problem can be given as

$$\max_{\boldsymbol{a}} \{\boldsymbol{Y}^\top \boldsymbol{a} | \boldsymbol{X}^\top \boldsymbol{a} = (1-\tau)\boldsymbol{X}^\top \mathbf{1}_n,\ \boldsymbol{a} \in [0,\ 1]^n\}, \tag{2.35}$$

where $[0,\ 1]^n$ is the n-dimentional unit cube. The dual solution $\hat{\boldsymbol{a}}(\tau)$ reduces to one sample problem of Hájek and Šidák (1967) rankscores process when design matrix $\boldsymbol{X}$

takes the simple form of an n vector of ones ($\boldsymbol{X} = \mathbf{1}_n$). As simple quantile regression problem produces sample quantiles, the dual problem generates the order statistics, or more precisely the *ranks* of the observations. For example, $\hat{\boldsymbol{a}}_i(\tau) = 0$ when $(\boldsymbol{x}_i,\ y_i)$ is below the fitted quantile line $\mathcal{Q} = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}(\tau)$, and $\hat{\boldsymbol{a}}_i(\tau) = 1$ when $y_i > \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$. This implies that for $y_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$, where $\tau \in (0,\ 1)$, $\hat{\boldsymbol{a}}_i(\tau)$ lies somewhere between zero and one. So just as the primal problem *sorts* the sample observations, the dual *ranks* the observations. If we would change τ from 0 to 1, the solution of (2.35) can be written as $\hat{\boldsymbol{a}}(\tau) = \big(\hat{a}_1(\tau),\ \hat{a}_2(\tau), \dots\ , \hat{a}_n(\tau)\big)^\top$ and we refer to it as *a rank score process*.

Fundamental theory of the rank based inference can be found in Hájek and Šidák (1967), Gutenbrunner and Jurečková (1992) and Gutenbrunner, Jurečková, Koenker and Portnoy (1993).

2.6.1 Rank-Score Tests

The Rank-Scores can be used as a basis for testing the general linear hypothesis. In particular rank-score tests in the regression setting can be seen as being broadly equivalent to the Lagrange multiplier (LM) approach to testing the general linear hypothesis. Within a general linear modelling framework the LM test is asymptotically equivalent to the Wald test.

The τ^{th} quantile rank-score test applies the τ^{th} quantile function , $\varphi_\tau(t) = \tau - I(t < \tau)$, on the $n \times 1$ vector of dual linear programming solutions. Those solutions are associated with estimating the reduced parameter model corresponding to constrains set by the null hypothesis on the full parameter model. Recall that the τ^{th} regression quantile ($0 \leqslant \tau \leqslant 1$) for the heteroscedastic linear-location model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{U}$ is defined as $\mathcal{Q}_Y(\tau|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta} + F^{-1}(\tau)\boldsymbol{\gamma}$. In this model, $\boldsymbol{Y} = (y_1, y_2 \dots y_n)^\top$ is a vector of dependent responses, $\boldsymbol{X}$ is a $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the $p \times 1$ vectors of unknown regression and scale parameters respectively. $\boldsymbol{\Gamma}$ is a $n \times n$ matrix where the n diagonal elements are the n corresponding ordered elements of the $n \times 1$ vector $\boldsymbol{X}\boldsymbol{\gamma}\big(diag(X\gamma)\big)$ and $\boldsymbol{U}$ is a $n \times 1$ vector of random iid

errors from distribution function F.

Let us consider the partitioned linear quantile regression model as a full parameter model

$$Y = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\Gamma}\boldsymbol{U},$$

constructed by partitioning $\boldsymbol{X} = [\boldsymbol{X}_1 \vdots \boldsymbol{X}_2]$, where $\boldsymbol{X}_1$ is $n \times (p-q)$ and $\boldsymbol{X}_2$ is $n \times q$. In this case $\boldsymbol{\beta}(\tau) = \big(\boldsymbol{\beta}_1(\tau), \boldsymbol{\beta}_2(\tau)\big)$, where $\boldsymbol{\beta}_1(\tau)$ is a $(p-q) \times 1$ vector of unknown nuisance parameters under the null hypothesis, and $\boldsymbol{\beta}_2(\tau)$ is a $q \times 1$ vector of parameters specified by the null hypothesis

$$H_0 \colon \boldsymbol{\beta}_2(\tau) = \mathbf{0},$$

which is tested against the alternative hypothesis $H_1 \colon \boldsymbol{\beta}_2(\tau) \neq \mathbf{0}$. The $n \times 1$ vector of rank scores $\boldsymbol{R}(\tau) = \hat{\boldsymbol{a}}(\tau) - (1-\tau)\mathbf{1}$, where $\mathbf{1}$ denotes an $n \times 1$ vector of 1's, is regressed on the design matrix (Cade et al., 2006).

Gutenbrunner et al. (1993) propose to use the rank score test statistics:

$$T_\tau = \frac{\boldsymbol{S}^\top(\tau)\boldsymbol{Q}^{-1}\boldsymbol{S}(\tau)}{\tau(1-\tau)}, \tag{2.36}$$

where

$$\begin{aligned} \boldsymbol{S}(\tau) &= n^{-1/2}(\boldsymbol{X}_2 - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2)^\top\boldsymbol{R}(\tau) \quad \text{and} \\ \boldsymbol{Q} &= n^{-1}\boldsymbol{X}_2^\top(\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top)\boldsymbol{X}_2. \end{aligned}$$

Basing the proof on a linear quantile regression model with iid random errors they show that, under H_0, the asymptotic distribution of T_τ is central χ^2 with q degrees of freedom.

Koenker and Machado (1999) also consider the rank score test for location-scale linear regression models, concluding that the test performs '*quite well*'. Wei (2004) extends the aplication of the rank-score test and shows how to use it for semiparametric longitudinal quantile regression models.

An important attraction of the rank score test statistic T_τ is that it does not require estimation of the nuisance, sparsity parameter. The test statistic T_τ can be

applied either to examining a global effect of the covariates, $\boldsymbol{X}_2$, on the response $\boldsymbol{Y}$ or by chosing to focus only on a single τ as a purely local, quantile-specific test.

However, one of the most important applications of the classical theory of rank tests involves the construction of confidence intervals for individual parameters of the quantile regression model. Considering the same model as earlier, but this time testing the hypothesis $H_0 : \boldsymbol{\beta}_2(\tau) = \boldsymbol{\eta}$ for a fixed quantile τ, integrating $\hat{a}_i(\tau)$ with respect to the τ-quantile score function

$$\varphi_\tau(t) = \tau - I(t < \tau),$$

general form of the rank-score test statistic can be given by

$$T_n = \frac{\boldsymbol{S}_n^\top \boldsymbol{Q}_n^{-1} \boldsymbol{S}_n}{\tau(1-\tau)}. \tag{2.37}$$

Under the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\eta}$ (often $\boldsymbol{\eta} = \mathbf{0}$)

$$\boldsymbol{S}_n(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \boldsymbol{X}_2^\top \hat{\boldsymbol{b}}_n(\boldsymbol{\eta}) \to \mathcal{N}\big(0, \tau(1-\tau)\boldsymbol{\Omega}_n\big),$$

where $\boldsymbol{\Omega}_n = \frac{1}{n}\boldsymbol{X}_2^\top \big(\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\big)\boldsymbol{X}_2$. We can calculate

$$T_n(\boldsymbol{\eta}) = \frac{1}{\sqrt{\tau(1-\tau)}} \boldsymbol{S}_n(\boldsymbol{\eta}) \boldsymbol{\Omega}_n^{-1/2}$$

and reject H_0 if $|T_n(\boldsymbol{\eta})| > \Phi^{-1}(1-\alpha/2)$.

The main advantage of this approach is that it inherits the scale invariance of the test statistic T_n and ultimately avoids the problem of explicit estimation of the sparsity function, that on the other hand, could be viewed as a disadvantage as it does not allow for estimation of the variance-covariance matrix. In practice, this method is convenient only when estimating confidence intervals of one-dimentional parameters. Even for component-wise confidence intervals, the computational complexity of the rank-score method grows with the size of the data sets (Kocherginsky et al., 2005). An alternative route to the inference of regression quantiles is offered by resampling methods, in which the construction of the confidence intervals based on the quantile regression estimator can be greatly simplified.

Chapter 3

Resampling Schemes When the Data are Normal

3.1 Introduction

As we have seen from the previous subsections, the asymptotic precision of the quantile estimates depend upon the reciprocal of the error density, a quantity we refer to as *sparsity*, where estimation proves to be a difficult task. Apart from being able to use rank-score method to overcome this difficulty, when the inferental information of a quantile of interest is required, we can also use a bootstrapping approach.

Bootstrapping is a computationally intensive, nonparametric technique that makes probability-based inference about a population characteristic, Θ, based on an estimator, $\hat{\Theta}$, using a sample drawn from a population. Bootstrapping relies on analogy between the sample and the population from which the sample was drawn by treating the sample as if it is a population. The data is resampled with replacement many times in order to obtain an empirical estimate of the sampling distribution of the statistic of interest Θ. Thus, the bootstrapping enables us to make inference without having to make distributional assumptions (see Efron, 1979).

There are several possible ways of implementing the bootstrap technique. Efron (1982, pg. 35) proposed the residual bootstrap for a nonlinear median regression

problem. This idea has been adopted for quantile regression setting, and its' implementation has been illustrated by several authors, e.g. Buchinsky (1994) and Hahn (1995).

By firstly fitting a quantile regression model and then obtaining the residuals

$$u_i(\tau) = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau) \tag{3.1}$$

we get a set of residuals $\{u_1(\tau), \ldots, \ u_n(\tau)\}$ from which a bootstrap resample $\{u^*_{b,1}(\tau), \ldots, \ u^*_{b,n}(\tau)\}$ is drawn with replacement. A bootstrapped vector of the response variable for this resample is generated by adding the resampled vector of residuals to the vector of fitted response values, $\hat{\boldsymbol{y}} = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}(\tau)$, from the sample:

$$\boldsymbol{y}^*_b = \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}^*_b(\tau). \tag{3.2}$$

We use these bootstrapped responses $y^*_{b,i}$ to estimate quantile regression coefficient by computing

$$\hat{\boldsymbol{\beta}}^*_b(\tau) = \min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_\tau(y^*_{b,i} - \boldsymbol{x}_i^\top \boldsymbol{\beta}), \tag{3.3}$$

where $\rho_\tau(u) = u(\tau \cdot 1_{[u \geqslant 1]} - (1-\tau) \cdot 1_{[u<0]})$. This procedure, from the residual resample to the estimation of $\hat{\boldsymbol{\beta}}^*_b(\tau)$, is repeated B times, giving us a $B \times p$ matrix in which $\hat{\boldsymbol{\beta}}^*_b(\tau)$ is a $p \times 1$ vector of quantile regression coefficients. Each column in this matrix of bootstrapped regression coefficients can be converted into an estimate of the sampling distribution of $\hat{\beta}_j(\tau), \ j = , 1 \ldots, \ p$, by placing probability of $1/B$ on each value of $\hat{\boldsymbol{\beta}}^*_b(\tau)$ for a given parameter β_j.

We have seen, using Theorem 2.3.1 that the sample τ-quantile ξ_τ converges weakly to $\mathcal{N}\big(0, \tau(1-\tau)/f(0)^2\big)$ and, by approximating this distribution, we are able to construct confidence intervals for $\boldsymbol{\beta}(\tau)$. It is shown (for example see Fitzenberger (1997)) that the distribution of the sample τ-quantile of the resampled residuals $\{u^*_{b,1}(\tau), \ldots, \ u^*_{b,n}(\tau)\}$ is a good approximation of $\mathcal{N}\big(0, \tau(1-\tau)/f(0)^2\big)$.

Assuming the idd error condition when applying the residual bootstrap, De Angelis et al. (1993) show that if we let $\boldsymbol{z} = (z_1, \ldots, z_p)^\top \in \mathbb{R}$, the bootstrap estimator of distribution G of $\hat{\boldsymbol{\beta}}$

$$G(\boldsymbol{z}) = P\big\{\sqrt{n}\big(\hat{\beta}_j(\tau) - \beta_{0j}(\tau)\big) \leqslant z_j, \ 1 \leqslant j \leqslant p\big\} \tag{3.4}$$

conditional on the initial sample $\mathcal{X} = \{(\boldsymbol{x}_i, Y_i) \colon i = 1, \ldots, n\}$, is given by

$$\hat{G}(\boldsymbol{z}) = P\big\{\sqrt{n}\big(\hat{\beta}_j^*(\tau) - \hat{\beta}_j(\tau)\big) \leqslant z_j,\ 1 \leqslant j \leqslant p | \mathcal{X}\big\}, \tag{3.5}$$

and converges to the limiting distribution of $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_b^*(\tau) - \hat{\boldsymbol{\beta}}(\tau)\big)$. Additionally, they also show that this approximation is of order $O(n^{-1/4})$ as $n \to \infty$.

Having $\hat{\boldsymbol{\beta}}_1^*(\tau), \ldots,\ \hat{\boldsymbol{\beta}}_B^*(\tau)$ and assuming the iid error, under the condition of normality given by Theorem 2.3.2, we can estimate the asymptotic variance-covariance matrix $\omega^2 \boldsymbol{Q}_0^{-1}$ of $\hat{\boldsymbol{\beta}}(\tau)$, where $\omega^2 = \tau(1-\tau)/f^2(F^{-1}(\tau))$ and $\boldsymbol{Q}_0^{-1}$ is $n^{-1}\boldsymbol{X}^\top \boldsymbol{X}$, by the bootstrap population variance-covariance matrix $\hat{\boldsymbol{\Sigma}}^*(\tau) = \mathrm{var}^* \sqrt{n}\big(\hat{\boldsymbol{\beta}}_b^*(\tau) - \hat{\boldsymbol{\beta}}(\tau)\big)$, which is a bootstrapped variance of $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_b^*(\tau) - \hat{\boldsymbol{\beta}}(\tau)\big)$, conditional on the initial sample. We compute an approximation to $\hat{\boldsymbol{\Sigma}}^*(\tau)$ by

$$\hat{\boldsymbol{\Sigma}}(\tau) = \frac{n}{B} \sum_{b=1}^{B} \big(\hat{\boldsymbol{\beta}}_b^*(\tau) - \overline{\hat{\boldsymbol{\beta}}}(\tau)\big)\big(\hat{\boldsymbol{\beta}}_b^*(\tau) - \overline{\hat{\boldsymbol{\beta}}}(\tau)\big)^\top, \tag{3.6}$$

where $\overline{\hat{\boldsymbol{\beta}}}(\tau) = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_b^*$ (Gonçalves and White, 2005). This is a consistent estimator of $\boldsymbol{\Sigma}(\tau)$ only under the assumption of independence. In the case when the independent assumption does not hold, the resampling scheme disrupts any relationship that might exist between the residual vector $\boldsymbol{U}(\tau)$ and $\boldsymbol{X}$ making this approach invalid.

De Angelis et al. (1993) compare this conventional strategy of direct estimation of the asymptotic covariance matrix with a non-paramertic approach in which the approximation is based on a kernel estimator of $f(F^{-1}(\tau))$

$$\hat{f}(F^{-1}(\tau)) = (nh)^{-1} \sum_{i=1}^{n} K(\hat{u}_i / h), \tag{3.7}$$

where K is a symmetric density function. They note that $f(F^{-1}(\tau))$ can be estimated to $O(n^{-2/5})$ accuracy. Subsequently, the error of approximation of the $\mathcal{N}(0,\ \hat{\omega}\boldsymbol{Q}_0^{-1})$ estimation is also of the same order $O(n^{-2/5})$. This opens up a question of whether the residual bootstrap approximation can be improved by some further smoothing technique. De Angelis et al. (1993) suggest further investigation by applying some of the smoothing methods for kernel density estimation proposed by Silverman (1986).

3.2 Bootstrapping Methods for Quantile Regression

The residual bootstrap is of limited practical interest for quantile regression, as in practice we can hardly expect to have a location-shift problem which assumes iid errors (Koenker, 2005, pg.107). In the simplest independent, but not identically distributed setting, the $(x,\ y)$ pair method of the bootstrap provides an effective alternative which can accommodate some forms of heteroscedasticity. Instead of drawing bootstrap samples from the empirical distribution of the residuals, samples of the (x_i, y_i) pairs are drawn from the the joint empirical distribution of the sample. Consequently, $(x_i^*,\ y_i^*)$ is drawn with replacement from the n pairs $\{(x_i,\ y_i)\colon\ i = 1, \ldots, n\}$ of the original sample with the probability of $1/n$.

One of the practical aspects that needs to be considered when implementing any of the bootstrapping techniques for the quantile regression models is the choice of the number of replications B. Number of bootstrap repetitions B for bootstrap standard error, confidence intervals, confidence regions, hypothesis testing, $p-values$ and bias corrections has been addressed by Andrews and Buchinsky (2000). They provide a comprehensive study which, amongst others, includes bootstraps for regression models based on bootstrapping residuals. They determine a formula for how large B needs to be to attain a desired level of accuracy based on the asymptotic approximation. In earlier work, Buchinski (1995) reports on an extensive Monte Carlo study in which he compares several bootstrapping methods. In this study, Buchinsky concludes that the $(x,\ y)$-pair performs well, but a more valuable feature of his experiments was comparisons of the bootstrap samples, m, and the size of the original sample, n. He suggest's that $m < n$ produces more accurate confidence intervals than when $m = n$ when applying residual bootstrap, wheres the performance is about the same for $(x,\ y)$ bootstrap. This finding is of interest, since when using the bootstrapping to approximate the sample median Sakov and Bickel (2000) show that $m < n$.

Another approach to bootstrapping for quantile regression is offered by the realisation that, rather than bootstrapping $(x_i,\ y_i)$ pairs, we can instead bootstrap the quantile regression gradient condition. Unlike these classical bootstrapping methods

that could be implemented to compute confidence intervals for the regression quantile, Parzen, Wei and Ying (1994) consider the estimation function of the τ^{th} quantile

$$S(\boldsymbol{b}) = -n^{1/2} \sum_{i=1}^{n} \boldsymbol{x}_i \big(\tau - I(y_i \leqslant \boldsymbol{x}_i^\top \boldsymbol{b})\big),$$

observing it as a pivotal statistic for the true τ^{th} quantile regression parameter when evaluated at $\boldsymbol{b} = \boldsymbol{\beta}(\tau)$, which makes its distribution independent of the parameter $\boldsymbol{\beta}(\tau)$. Replacing $\boldsymbol{\beta}(\tau)$ with an estimate yields an asymptotically pivotal quantity. The distribution of $S(\boldsymbol{b})$ may be generated exactly by the random vector $\boldsymbol{U}$,

$$\boldsymbol{U} = -n^{1/2} \sum_{i=1}^{n} \boldsymbol{x}_i(\tau - \xi_i)$$

which is a weighted sum of independent, re-centered Bernoulli variables ξ_i, taking the value 1 with probability τ and 0 with probability $1 - \tau$.

For a given realisation $\boldsymbol{u}$ from $\boldsymbol{U}$ we can consider a problem of finding

$$\boldsymbol{\beta}^*_{\boldsymbol{U}} = \{\boldsymbol{b} \colon S(\boldsymbol{b}) = \boldsymbol{u}\}$$

by solving the increasing quantile regression problem

$$\boldsymbol{\beta}^*_{\boldsymbol{U}} = \min_{\boldsymbol{b} \in \mathbb{R}} \sum \rho_\tau(y_i - \boldsymbol{x}_i \boldsymbol{b}) + \rho_\tau(\zeta - \sqrt{n} \boldsymbol{u}_b^\top \boldsymbol{b}/\tau),$$

in which ζ is chosen to be sufficiently large to insure that it exceeds $\sqrt{n}\boldsymbol{u}^\top \boldsymbol{b}/\tau$. This ensures that the contribution of the last term to the sub-gradient is

$$\sqrt{n}\boldsymbol{u} = \sum \boldsymbol{x}_i(\tau - \xi_i)$$

and therefor, that $S(\boldsymbol{b}) = \boldsymbol{u}$ is satisfied. Parzen et al. (1994) prove that, for the large sample sizes, the distribution of $\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)$ can be approximated by a conditional distribution of $\hat{\boldsymbol{\beta}}^*_{\boldsymbol{U}} - \hat{\boldsymbol{\beta}}(\tau)$, where $\hat{\boldsymbol{\beta}}^*_{\boldsymbol{U}}$ solves the increasing quantile regression problem $S(\boldsymbol{b}) = \boldsymbol{u}$ with $n + 1$ observations and $x_{(n+1)} = -n^{1/2}u/\tau$ and $y_{(n+1)}$ is an extreamly large number for a given realisation $\boldsymbol{u}_b$ from $\boldsymbol{U}$.

Parzen, Wei and Ying's approach achieves a robustness that accommodates certain forms of heteroscedasticityby by benefiting from the asymptotically pivotal role of the quantile regression "gradiant condition". Although this bootstrap method has

a simple concept, Chen and Wei (2005) point out, that when using it for relatively large data sets, in particular for high-dimentional data sets, the method might take longer time when compared with other, alternative methods. This form of bootstrapping for quantile regression has been implemented in R's `quantreg` package: `boot.rq.pwy`.

Markov chain marginal bootstrap (MCMB) is another general resampling method for constructing confidence intervals of certain parametric models and for a wide class of M estimators of linear regression, which was proposed by He and Hu (2002). Unlike the classical bootstrap approach, the MCMB method distinguishes itself in two key aspects: it involves solving only one-dimensional equations for parameters of any dimension and produces a Markov chain rather than a conditionally independent sequence. In terms of its application to the quantile regression problem this means that the MCMB solves p one-dimentional equations instead of p-dimensional equations as previous methods do.

The MCMB approach (Kocherginsky et al., 2005) requires the linearity only of the τ^{th} quantile for one given level of τ, not for other percentiles. Let $x_{i,j}$ be the j^{th} component of x_i, $x_{i,(j-)}$ and $x_{i,(j+)}$ as the first $j-1$ and the last $p-j$ components of x_i, respectively. Subsequently, we can write $\boldsymbol{x}_i^\top \boldsymbol{\beta} = x_{i,j}\beta_j + x_{i,(j-)}^\top \beta_{(j-)} + x_{i,(j+)}^\top \beta_{(j+)}$ for any $1 \leqslant j \leqslant p$. Let $\psi(\tau)$ be the derivative of the score function $\rho(\tau)$

$$n^{-1} \sum_{i=1}^{n} \psi(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i = 0. \tag{3.8}$$

Let $\boldsymbol{z}_i = \psi(\tau)(u_i)\boldsymbol{x}_i - \bar{\boldsymbol{z}}$, where $\bar{\boldsymbol{z}} = n^{-1} \sum_{i=1}^{n} \psi(\tau)(u_i)\boldsymbol{x}_i$, and $u_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$ be the residual. The MCMB algorithm starts from the quantile estimate $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}(\tau)$ with steps $k = 0$ and iterates through the following steps:

1. $k \leftarrow k + 1$

2. For each j starting from 1 to p ($j \in [1,\ p]$), $\{z_1, \ldots, z_n\}$ are drawn with replacement to obtain $\{z_1^{k,j}, \ldots, z_n^{k,j}\}$ in order to solve $\beta_j^{(k)}$ as the root to

$$\sum_{i=1}^{n} \psi(\tau)\big(y_i - x_{i,(j-)}^\top \beta_{(j-)}^{(k)} - x_{i,j}\beta_j^{(k)} - x_{i,(j+)}^\top \beta_{(j+)}^{(k-1)}\big) x_{i,j} = \sum_{i=1}^{n} z_i^{k,j}. \tag{3.9}$$

3. Steps 1 and 2 are repeated until a pre-specified number of replications K is reached.

The key to the MCMB algorithm is in the step 2. An independent sample $\{z_1^{k,j}, \ldots, z_n^{k,j}\}$ has to be drawn for each j at the k^{th} step[1]. Using equation (3.9) we are solving for $\beta_j^{(k)}$ by using the most recent values of other parameters. This way we would obtain a Markov chain sequence $\beta^{(1)}, \ldots, \beta^{(K)}$. He and Hu (2002) also propose a way for approximating the joint distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. By solving p one dimensional equations, rather than p dimensional system, a Markov chain for each component of $\boldsymbol{\beta}$ is generated. The variance-covariance matrix of the distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can then be aproximated by the second moment of the Markov chain, under the assumption that the average $n^{-1}\sum_{i=1}^{n} \psi(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})\boldsymbol{x}_i$ converges, as $n \to \infty$, to a continuously differentiable function $U(\beta)$ as a gradient of some strictly convex function of β, with continuous second order derivative matrix. In addition, they point out that the MCMB approach, unlike the traditional (x_i, y_i)-paired bootstrap, fails to be asymptotically correct in the models with a large p and in general heteroscedastic models, unless a correct likelihood is specified. However, in their simulation studies that are based on models with small p, they illustrate how this approach is not very sensitive to certain types of deviations from iid errors. This method proves to be highly robust against many forms of moderate heteroscedasticity.

Chen and Wei (2005) in their simulation study test the performance of the confidence interval methods for the large p quantile regression model. They demonstrate that rank-score and MCMB methods show good performance with large p, which reinforces the recommendation by Kocherginskey et al. (2005) which suggests, that the MCMB method can provide good results in estimating variance-covariance matrix and construction of SD-confidence intervals for moderately large problems. However, Kocherginskey et al. (2005) in their Recommendation 1, point out that in order to have reliable inference, for a p-dimensional regression quantile, it is desirable to have $n \min\{\tau, 1-\tau\} > 5p$, for a sample size of n. This bootstrapping technique for the quantile regression model is available from `quantreg` package and the package

[1]Note, the left side of the equation (3.9) is a monotone step function and the root to (3.9) is interpreted as the point of sign change.

`rqmcmb2` developed by Masha Kocherginsky and Xuming He for `R`.

Koenker has implemented another resampling method for quantile regression model in his `quantreg` package. It is a generalised bootstrap technique for estimators obtained by minimising functions that are convex in the parameter using unit exponential weights, which was developed by Bose and Chatterjee (2003).

Let $X_1, \ldots, X_m$ be m independent and identically distributed copies of X-valued random variable and let $f(b, x)$ be a real measurable function defined for $b \in \mathbb{R}^p$, $p \geqslant 1$, $x \in X^m$, $m \geqslant 1$. We can consider

$$\mathcal{Q}(b) = Ef(b, X_1, \ldots, X_m). \tag{3.10}$$

Let us further assume that there is a unique $b_* \in \mathbb{R}^p$ such that

$$\mathcal{Q}(b_*) = \min_b \mathcal{Q}(b), \tag{3.11}$$

b_* is the unknown parameter to be estimated from the data. Under the assumption that $X_1, \ldots, X_n$ is an iid sample, we can consider a sample analogue to (3.10) of the form

$$\mathcal{Q}_n(b) = \begin{pmatrix} n \\ m \end{pmatrix}^{-1} \sum_{1 \leqslant i_1 < \cdots i_m \leqslant n} f(b, X_{i_1}, X_{i_2}, \ldots, X_{i_m}) \tag{3.12}$$

and minimise $\mathcal{Q}_n(b)$, where b_n is such that

$$\mathcal{Q}(b_n) = \min_b \mathcal{Q}_n(b). \tag{3.13}$$

The statistic b_n is the M_m estimator introduced by Huber (1964). Bose and Chatterjee (2003) suggest further that, since in practice most criteria functions are convex, we can assume that $f(b, x)$ is convex in b. Three most common examples of such functions would be the mean ($m = 1$, $p = 1$ and $f(b, x) = (b - x)^2 - x^2$); the median ($m = 1$, $p = 1$ and $f(b, x) = |b - x| - |x|$), and the sample variance ($m = 2$, $p = 2$ and $f(b, x_1, x_2) = (b - (x_1 - x_2)^2/2)^2 - (x_1 - x_2)^4/4$).

Bose and Chatterjee's (2003) bootstrapping approach for approximating the distribution of b_n falls into the class of "*weighted bootstrap*". For every $n \geqslant 1$ and every $i_1, i_2, \ldots, i_m$ distinct $i_j \in \{1, 2, \ldots, n\}$; $j = 1, \ldots, m$, let $\{w_{n:\, i_1, i_2, \ldots, i_m}\}$ be real-value non-negative random variables independent of $\{X_i\}$, representing the "*bootstrap*

weights". In this case the bootstrap equivalent of b_n will be obtained by minimising

$$\mathcal{Q}_n(b) = \left(\begin{array}{c} n \\ m \end{array} \right)^{-1} \sum_{1 \leqslant i_1 < \cdots i_m \leqslant n} w_{n:\, i_1, i_2, \ldots, i_m} f(b, X_{i_1}, X_{i_2}, \ldots, X_{i_m}) \tag{3.14}$$

As already pointed out $\{b_{nB}\}$ is chosen to satisfy

$$\mathcal{Q}_{nB}(b_{nB}) = \min_b \mathcal{Q}_{nB}(b).$$

Although the method has been adopted for quantile regression models and made available for easy implementation within the `quantreg` package in `R`: `boot.rq.wxy`, there are not many studies that include this technique in their comparisons via simulations. However, Koenker (2005), in Chapter 3 of his book, provides a study of this kind in which *wxy* method is competing ahainst *xy* and *MCMB* when used for the models with iid errors and models with not identically distributed (nid) errors. Koenker reports the results for confidence intervals and for both type of problems *wxy* outperforms the other two approaches in terms of the coverage, but provides slightly wider length. In our study we also include this bootstrapping technique as a competing method (see the results presented in the following chapters).

3.3 Smooth Bootstrapping Using Conditional Variance Modelling

In our approach for estimating confidence intervals for quantile regression functions we propose to use a smoothed bootstrap technique. When using the standard bootstrap the empirical distribution F_n is a discrete distribution and therefore samples constructed from F_n in the bootstrap simulations will have some unusual properties (Silverman and Young, 1987). That is, all the values taken by the members of the bootstrap samples will be drawn from the original sample values, thus nearly every sample will contain repeated values. The smooth bootstrap suggested by Efron (1979) is a modification to the standard bootstrapping procedure, designed to avoid samples with these properties. The basic idea of the smooth bootstrap is to perform the repeated sampling not from F_n itself, but from a smoothed version $\hat{F}$ of F_n.

Suppose we have a set of observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn from a distribution with an unknown density f. Those observations can be used to construct a nonparametric estimate $\hat{f}$ of the density f, after which as many independent realisations as required can be drawn from $\hat{f}$ to make up the bootstrap samples. We can construct $\hat{f}$ by the simple kernal method with kernel K and window width h, also called the *smoothing parameter* or *bandwidth*:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{3.15}$$

Assuring that the kernel K is non-negative and that satisfies the following condition

$$\int_{-\infty}^{\infty} K(x)dx = 1,$$

which is a probability density function, then $\hat{f}$ will itself be a probability density. Additionally, all the continuity and differentiability properties of the kernel K will be passed onto $\hat{f}$. If, for example, K is the normal density function, in that case $\hat{f}$ will be a smooth curve having derivatives of all orders (see Silverman, 1986, Chapter 2.4). In order to ensure that the weights are symmetric about the observations, another condition has to be satisfied

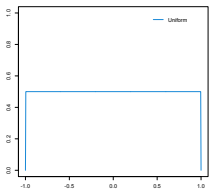
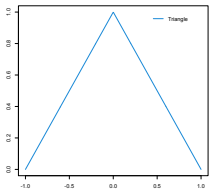
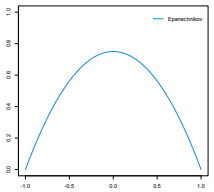
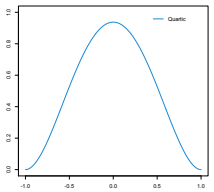
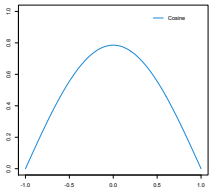
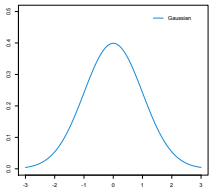
$$K(-x) = K(x), \quad \text{for all } x,$$

making sure that K is a symmetric probability density function. Kernel function K, is typically chosen to be a smooth unimodal function with a peak at 0. Table 3.1 provides a list of commonly used kernel density functions together with their illustrations.

Independent realisations from $\hat{f}$ are easy to find when $\hat{f}$ is constructed by a kernel method, as long as a non-negative, symmetric kernel is used. In fact, finding $\hat{f}$ explicitly is not even necessary in the simulation procedure. It is easy to simulate from $\hat{f}$ by sampling with replacement from the original data and transforming each sample point appropriately. Sometimes, we might be in the situation when it is desirable to simulate not explicitly from $\hat{f}$, but from a version transformed to have the same mean and variance as the observed data. Silverman (1986, Section 6.4.1)

proposes that realisations Y from $\hat{f}$ can be generated, for a univariate case in which $\hat{f}$ has been constructed by a standard kernel method with kernel K and bandwidth

Table 3.1: Kernel functions in common use

Kernel function, $K(t)$		
Uniform	$K(t) = \frac{1}{2} I_{\{\lvert t\rvert \leqslant 1\}}$	
Triangular	$K(t) = (1 - \lvert t\rvert) I_{\{\lvert t\rvert \leqslant 1\}}$	
Epanechnikov	$K(t) = \frac{3}{4}(1 - t^2) I_{\{\lvert t\rvert \leqslant 1\}}$	
Quartic	$K(t) = \frac{15}{16}(1 - t^4) I_{\{\lvert t\rvert \leqslant 1\}}$	
Cosine	$K(t) = \frac{\pi}{4} \cos(\frac{\pi}{2} t) I_{\{\lvert t\rvert \leqslant 1\}}$	
Gaussian	$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$	

h, as follows:

- *Step 1*: Choose I uniformly with replacement from $\{1, \ldots, n\}$;
- *Step 2*: Generate ϵ to have probability density function K;
- *Step 3*: Set $Y = X_I + h\epsilon$.

In *Step 2* it is necessary to generate a random observation from the kernel K. This algorithm is easy to apply and it can be repeated as often as necessary to give independent realisations Y_j from $\hat{f}$.

In the case when the realisations Y are transformed to reflect the first and second moment properties observed in the sample $\{X_1, \ldots, X_n\}$, *Step 3* of the algorithm should be replaced by

$$\textit{Step 3'}: \quad Y = \bar{X} + (x - \bar{X} + h\epsilon)/(1 + h^2\sigma_K^2/\sigma_X^2)^{1/2}$$

where $\bar{X}$ and σ_X^2 are the sample mean and variance of $\{X_i\}$ and σ_K^2 is the variance of the kernel K.

Suppose that we have $\{y_1, \ldots, y_n\}$ a random sample of response variable Y, $\boldsymbol{x}_i \in \mathbb{R}^p$ is a covariate vector coresponding to the i^{th} observation y_i and assume that τ^{th} conditional quantile function is

$$\mathcal{Q}_y(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}(\tau) \tag{3.16}$$

for some parameter $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$. The asymptotic normality for the quantile estimate $\hat{\boldsymbol{\beta}}(\tau)$ is established under the assumption of iid error model, that is

$$u_i(\tau) = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau), \tag{3.17}$$

where u_i are iid variables with the τ^{th} quantile at 0 (see Section 2.3). By using this set of residuals $\{u_1(\tau), \ldots, u_n(\tau)\}$ to construct a non-parametric estimate $\hat{f}_{u(\tau)}$ of the $f_{u(\tau)}$, we can adapt Silverman's smooth bootstrap algorithm to estimate standard errors, variance-covariance matrix as well as confidence intervals of the quantile estimators $\boldsymbol{\beta}(\tau)$. In the cases when the errors are iid, we can pool the residuals

u_i's together and estimate their common density $\hat{f}_{u(\tau)}$ from which to sample. In the non-iid errors case, we can standardise them by using an estimate of the conditional variance function.

We can summarize the method as following:

i) Estimate the τ^{th} quantile function of interest: $\hat{\mathcal{Q}}_y(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$;

ii) Obtain the residuals $\{u_1(\tau), \ldots, \ u_n(\tau)\}$: $u_i(\tau) = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau)$;

iii) Using the estimate of the mean function of the residuals,
$\hat{E}\big[u_i(\tau|\boldsymbol{x}_i)\big] = \tilde{u}_i(\tau) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_u + \epsilon_u$,
construct a set of centered and squared residuals: $su_i(\tau) = \Big(u_i(\tau) - \tilde{u}_i(\tau)\Big)^2$;

iv) Estimate the conditional variance function of the residuals at each x_i,
i.e. $\hat{V}\big[u_i(\tau)\big] = \hat{E}\big[su_i(\tau)\big]$ and then standardise each residual by:
$stu_i(\tau) = u_i(\tau)/\sqrt{\hat{V}[u_i(\tau)]}$;

v) Construct the kernel density estimate of the standardised residuals:

$$\hat{f}_{u(\tau)}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - stu_i(\tau)}{h}\right),$$

where K is the kernel function at $h > 0$ is the smoothing parameter.

vi) Draw a sample of standardised residuals from $\hat{f}(stu)$ using Silverman' algorithm (1986, pg.143):

Step 1: Chose I uniformaly with replacement from $\{1, \ldots, \ n\}$;

Step 2: Generate ϵ to have probability density function K;

Step 3: Set $stu_i^* = stu_I + h\epsilon$, or in the case when the realisations stu_i are transformed to reflect the first and second moment properties observed in the sample $\{stu_1, \ldots, \ stu_n\}$, use
$stu_i^* = \mu_{stu_i} + (stu_i - \mu_{stu_i} + h\epsilon)/\sqrt{(1 + h^2\sigma_K^2/\sigma_{stu_i}^2)}$;

vii) Scale the standardised residuals to their original scale and construct a smooth bootstrap sample $y_i^* = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau) + stu_i^* \sqrt{\hat{V}[u_i(\tau)]}$;

vii) Re-fit the quantile regression model to the bootstrap data

$$\mathcal{Q}_{y_b^*}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_b^*(\tau)$$

to obtain a new set of parameter estimates $\hat{\boldsymbol{\beta}}_b^*(\tau)$.

This procedure can be used in both iid and nid cases. An alternative in the iid case is to simply pool all the residuals together and then use these pooled values to estimate their density in order to obtain a smoothed bootstrap sample.

3.3.1 Conditional Variance Modelling

The parameter estimate and related statistical inference for the linear regression models is commonly based on the assumption that the error terms are homoscedastic. As we have already pointed out, when dealing with the quantile regression problem in practice, this assumption usually is not guaranteed. Accurate estimation of the conditional variance functions of the error term in a heteroscedastic quantile regression model could be of significant importance for obtaining the quantile estimators $\boldsymbol{\beta}(\tau)$ and making valid statistical inferences.

If $Z_1, Z_2, \ldots, Z_n$ is a random sample from a standard normal distribution it is known that (Mood, Graybill and Boes, 1974):

1. $\bar{Z}$ has a normal distribution with mean 0 and variance $1/n$.
2. $\bar{Z}$ and $\sum_{i=1}^n (Z_i - \bar{Z})^2$ are independent.
3. $n\bar{Z}^2$ has a chi-square distribution with one degree of freedom.
4. $\sum_{i=1}^n (Z_i - \bar{Z})^2$ has a chi-square distribution with $n-1$ degrees of freedom.

We also note that a chi-square density function is a particular case of a gamma density function.

Definition 1. **Chi-square distribution** If X is a random variable with *pdf*

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} I_{\{x \geqslant 0\}},$$

then X is defined to have *chi-square distribution* with k degrees of freedom.

Definition 2. **Gamma distribution** If a random variable X has *pdf* given by

$$f(x;\nu,\lambda) = \frac{\lambda}{\Gamma(\nu)}(\lambda x)^{\nu-1}e^{-\lambda x}I_{(0,\infty)}(x),$$

where $0 < \nu < \infty$, $0 < \lambda < \infty$ and $0 < x < \infty$, then X is defined to have a *gamma distribution*, with

$$E[X] = \mu = \frac{\nu}{\lambda}, \quad var[X] = \frac{\nu}{\lambda^2} = \frac{\mu^2}{\nu}$$

and moment generating function

$$m_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\nu} \quad \text{for} \quad t < \lambda.$$

If we set the parameters ν and λ, of the gamma distribution to be equal to $k/2$ and $1/2$ respectively, then we get the chi-square distribution. Thus, if the X random variable has a chi-square distribution,

$$\begin{aligned} E[X] &= \frac{k/2}{1/2} = k, \\ var[X] &= \frac{k/2}{(1/2)^2} = 2k \quad \text{and} \\ m_X(t) &= \left[\frac{\frac{1}{2}}{\frac{1}{2} - t}\right]^{k/2} = \left[\frac{1}{1-2t}\right]^{k/2} \end{aligned}$$

at the same time, we can say that X also has gamma distribution.

Let us apply the above theory and assume that the τ^{th} quantile residuals given in (3.17) are from a standard normal distribution, ie. $\boldsymbol{U} \sim \mathcal{N}(0,1)$, then $\boldsymbol{U}^2 \sim \chi^2(1)$ and the moment-generating function is

$$m_{\boldsymbol{U}^2}(t) = \left(\frac{1}{1-2t}\right)^{1/2}. \tag{3.18}$$

Now, let assume that $\boldsymbol{U} = \sigma\boldsymbol{X}$, and accordingly $\boldsymbol{U} \sim \mathcal{N}(0,\sigma^2)$. That means that

$$\begin{aligned} m_{\boldsymbol{U}^2}(t) &= m_{\sigma^2\boldsymbol{X}^2}(t) \\ &= m_{\boldsymbol{X}}(\sigma^2 t) \\ &= \left(\frac{1}{1-2\sigma^2 t}\right)^{1/2} \\ &= \left(\frac{\frac{1}{2\sigma^2}}{\frac{1}{2\sigma^2} - t}\right)^{1/2}. \end{aligned}$$

Thus, $\boldsymbol{U}^2 \sim Gamma(\frac{1}{2\sigma^2}, \frac{1}{2})$, resulting in

$$E\big[\boldsymbol{U}^2\big] = \frac{1/2}{1/(2\sigma^2)} = \sigma^2. \tag{3.19}$$

This implies, that by modeling the conditional mean of the squared centered residuals we can estimate the conditional variance function of the τ^{th} quantile error term, and further it suggests that for normally distributed error we can parametrically estimate the conditional variance function using a Gamma Generalised Linear Model (GLM).

Generalised linear models introduced by Nelder and Wedderburn (1972) are an extension of classical linear models, thus the application of the linear models accommodates response variables with non-normal conditional distribution. A GLM is defined by specifying two components:

1) *the response* should be a member of the exponential family distribution

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{3.20}$$

thus

$$\ell = \log f(y|\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \tag{3.21}$$

for some functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$. The θ is called *canonical parameter* and represents the location, whilst ϕ represents the scale and is referred to as *dispersion parameter*. The dispersion, ϕ, represents the variability in y_i. The function $a(\phi)$ usually has the form ϕ/w, where w is known value called *prior weight*. The exponential family distributions have mean and variance:

$$E[Y] = \mu = b'(\theta) \tag{3.22}$$

$$Var[Y] = b''(\theta)a(\phi). \tag{3.23}$$

The mean is a function of θ, while the variance is a product of the location and scale functions. The function b'' is *the variance function* and describes the way in which the variance relates to the mean. Note that through equation (3.22) θ is a function of μ and the function b'' expressed in terms of μ rather than θ is denoted by $V(\mu)$;

2) *The link* function g describes how the mean of the response $E[Y] = \mu$ and a linear combination of the predictors $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{X}^\top \boldsymbol{\beta}$, are linked

$$\eta = g(\mu). \tag{3.24}$$

In other word, link function $g(\mu)$ transforms the expectation of the response to the linear predictor. In practice, any monotone continuous and differentiable function can be used. However, there are some mathematically and computationally convenient and common choices for the standard GLMs, known as *canonical links* for which $\eta = g(\mu) = \theta$ and $g\big(b'(\theta)\big) = \theta$.

The parameters in a GLM are estimated by the values that maximise the likelihood L or, equivalently, the log-likelihood, ℓ, of the observed responses. If the observed y_i's are independent then their likelihood function is

$$L(\mu_1, \mu_2, \ldots, \mu_n, \phi) = \prod_{i=1}^{n} f(y_i|\mu_i, \phi) = \prod_{i=1}^{n} L(\mu_i, \phi). \tag{3.25}$$

Hence, the log-likelihhod function is

$$\begin{aligned} \ell(\mu, \phi) &= \ell(\mu_1, \mu_2, \ldots, \mu_n, \phi) && (3.26) \\ &= \sum_{i=1}^{n} \log L(\mu_i, \phi) = \sum_{i=1}^{n} \ell(\mu, \phi). && (3.27) \end{aligned}$$

The resulting *maximum likelihood estimates* (MLE) of $\boldsymbol{\beta}$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)^\top$, are obtained by differentiating $\ell(\mu, \phi)$ with respect to β_j and equating to zero for all j, $j = 1, \ldots, k$.

Gamma GLM

The density of the gamma distribution is usually given by:

$$f(y) = \frac{1}{\Gamma(\nu)} \lambda^\nu y^{\nu-1} e^{-\lambda y}, \qquad y > 0,$$

where ν describes the shape and λ describes the scale of the distribution (Figure 3.1). Thus, if Y has a gamma distribution, with parameters ν and λ,

$$E[Y] = \frac{\nu}{\lambda}, \qquad Var[Y] = \frac{\nu}{\lambda^2}, \qquad and \qquad m_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^\nu \quad for \;\; t < \lambda.$$

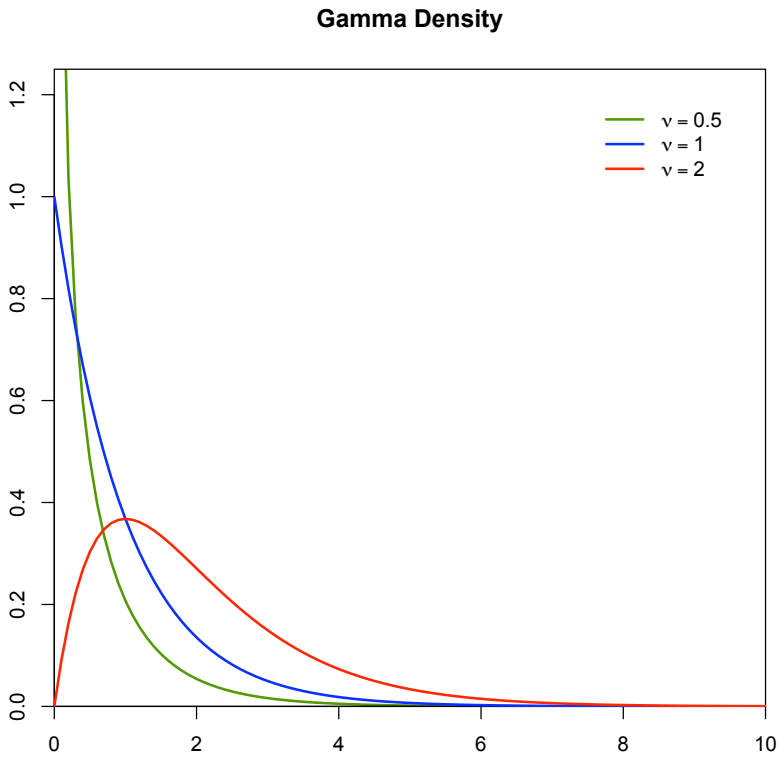


Figure 3.1: *Gamma Density: The graph plots 3 different gamma density functions. The green line presents gamma density with the shape parameter $\nu = 0.5$, from which we can see that the density is unbounded at zero. The blue line is gamma density with $\nu = 1$, which is the exponential density. The red line is gamma density with $\nu = 2$, where we can see a skewed distribution.*

For gamma family the *canonical parameter* is inverse, $\theta = -1/\mu$, so that the *canonical link* is

$$\eta = g(\mu) = -\mu^{-1} = -\frac{\nu}{\lambda}.$$

However, the minus is typically removed and we just use the inverse link. We also have

$$\begin{aligned} b(\theta) &= -\log(-\theta), \\ \phi &= \frac{1}{\nu}, \\ a(\phi) &= \phi \quad \text{and} \\ V(\mu) &= \mu^2. \end{aligned}$$

For gamma distribution, we can also use the *identity link*

$$\eta = g(\mu) = \mu = \frac{\nu}{\lambda} \quad \text{and}$$

and the *log link*

$$\eta = g(\mu) = \log \mu = \log \frac{\nu}{\lambda}.$$

Although, the canonical links in GLMs lead to desirable statistical properties of the model, that is for the canonical link the sufficient statistic is $\boldsymbol{X}^\top \boldsymbol{Y}$, where $\boldsymbol{X}$ is a $(n \times p)$ matrix and $\boldsymbol{Y}$ is a vector of size n, there is no reason why the linear combinations of the predictors in the model should be additive on the scale given by that link. The canonical links are convenient to use, but the convenience should not be the priority as a model selection criterion (Nelder and Wedderburn, 1972).

When considering the choice of GLMs, and any other models in general, we should define the range of the possibilities. The *null* model is the smallest model that conveys all the variation between the responses and their common mean value μ. It would be a model in which there is no relation between the predictors and the responses; a model in which we fit a common mean μ to all responses. Whilst, on the other hand, the *full* model is the most complex. The full model tries to explain the data exactly, and typically, we need to use n parameters for n data points, which could be achieved by using a sufficiently high-order polynomial. However, such a model would not tell us no more than the data itself and therefore it would be uninformative. Nevertheless, the full model can provide a baseline for measuring the discrepancy for an intermediate model with p parameters. The full model can provide us with a measure of how well a model fit the data, and therefore we can consider the difference between the log-likelihood for the full model, $\ell(y, \phi|y)$, and that for the model under consideration $\ell(\hat{\mu}, \phi|y)$, expressed as a likelihood ratio statistic

$$D(y, \mu) = 2\big(\ell(y, \phi|y) - \ell(\hat{\mu}, \phi|y)\big). \tag{3.28}$$

If we denote the estimates of the canonical parameters of the two models as $\hat{\theta} = \theta(\hat{\mu})$ and $\tilde{\theta} = \theta(y)$ respectively, and under the assumption that the observations are

independent and $a_i(\phi) = \phi/w_i$, the above statistic (3.28) can be written as

$$\sum_{i=1}^{n} 2w_i \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{\phi}. \tag{3.29}$$

This statistic can be simply written as $D^*(y, \hat{\mu})/\phi$, and we refer to $D(y, \hat{\mu})$ as *the deviance* of the current model, whilst $D^*(y, \hat{\mu})/\phi$ we call *the scale deviance.*

The form of the deviance for the gamma distribution is

$$D(y, \mu) = 2\Big[\sum_{i=1}^{n} \Big(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\Big(\frac{y_i}{\hat{\mu}_i}\Big)\Big)\Big] \tag{3.30}$$

and, therefore the scaled deviance for the gamma distribution is

$$D^*(y, \mu) = \nu D(y, \mu) = 2\nu\Big[\sum_{i=1}^{n} \Big(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\Big(\frac{y_i}{\hat{\mu}_i}\Big)\Big)\Big]. \tag{3.31}$$

3.4 Example 1: Applying the Smooth Bootstrap

Let us consider the following two models: a location-shift (homoscedastic) Model 1 (M1) and a location-scale shift (heteroscedastic) Model 2 (M2)

$$\begin{aligned} \text{Model 1} \quad &: \quad y_i = 2 + 5x_i + e_i \quad \text{and} \\ \text{Model 2} \quad &: \quad y_i = 2 + 5x_i + \sigma(x_i)e_i \end{aligned}$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$, and with $\{e_i\}$ iid from $\mathcal{N} \sim (0, 16)$ and $\sigma(x) = \sqrt{1 + 4x}$. We wish to estimate 75^{th} quantile, $\tau = 0.75$, thus 75^{th} quantile of the u_i variable is at 0 (see Figure 3.2). The fitted quantiles of the both models are given in Figure 3.3.

After fitting the 75^{th} quantile regression function we compute the residuals

$$u_i(\tau = 0.75) = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau = 0.75),$$

for which we estimate the conditional mean function

$$E\big[u_i(\tau = 0.75)|\boldsymbol{x}\big] = \tilde{u}_i(\tau = 0.75) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_u + \epsilon_u.$$

Having estimated the conditional mean function of the residuals we square the centered residuals

$$su_i(\tau = 0.75) = \Big(u_i(\tau = 0.75)) - \tilde{u}_i(\tau = 0.75)\Big)^2.$$

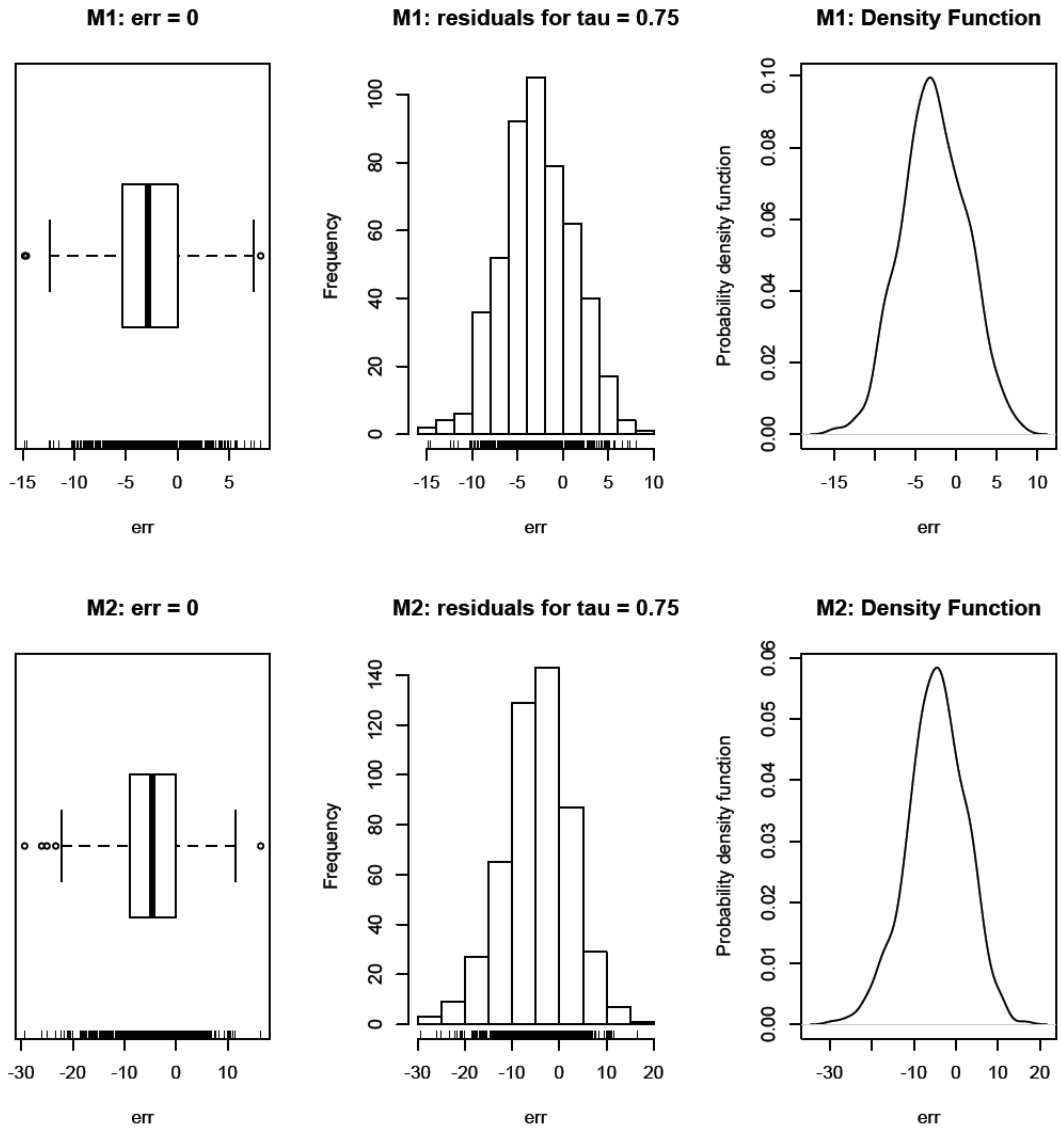


Figure 3.2: *Distribution of the* 75^{th} *Quantile Function Residuals: Box-Plots, Histograms and Density Functions illustrate how the* 75^{th} *quantile of the error term in the models is centered on 0, for* 75^{th} *quantile regression line of the data.*

Hence, by using (3.18) and (3.19), we know that conditional mean of those squared residuals is equal to the conditional variance and we can parametrically estimate the conditional variance function using a gamma GLM function in `R`

M1 : `V(u|x)=glm(`$su_i \sim$ `1, family=Gamma(link="inverse"))`

M2 : `V(u|x)=glm(`$su_i \sim$ `1 + x, family=Gamma(link="inverse")).`

Observing the scatter plot of the squared centered residuals of the location-scale shift

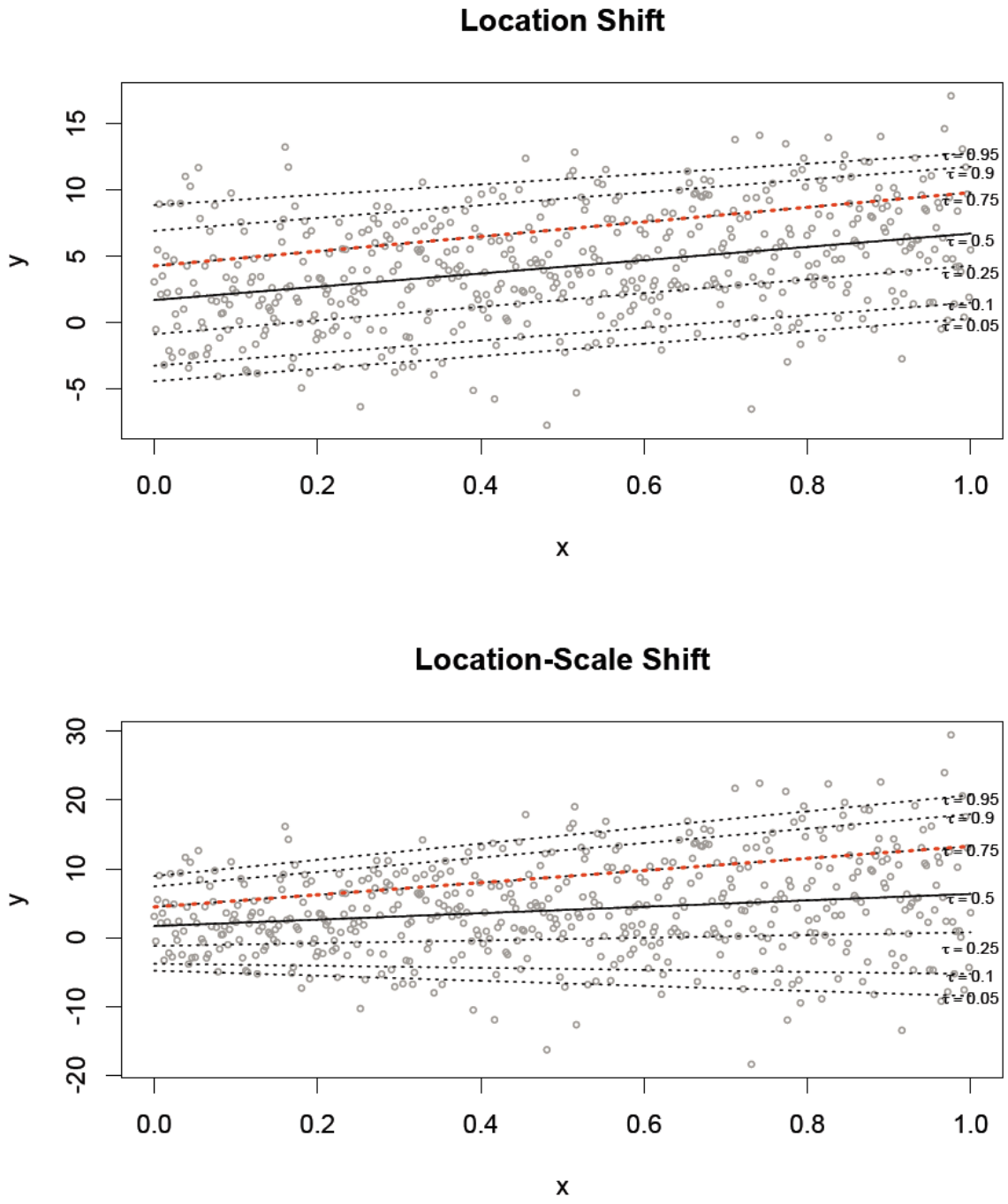


Figure 3.3: *Example 1 Data. The figures plot the simulated data for the Location and the Location-Scale Shift Models given in the Example 1. The data consists of 500 observations. Superimposed on the plots are seven estimated quantile regression lines corresponding to the quantiles* $\tau \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$.

model (Model 2, Figure 3.4), we can notice that the standard deviation is increasing linearly with the response $s\hat{u}_i$, suggesting that the coefficient of variation, $c_v = \sigma/\mu$,

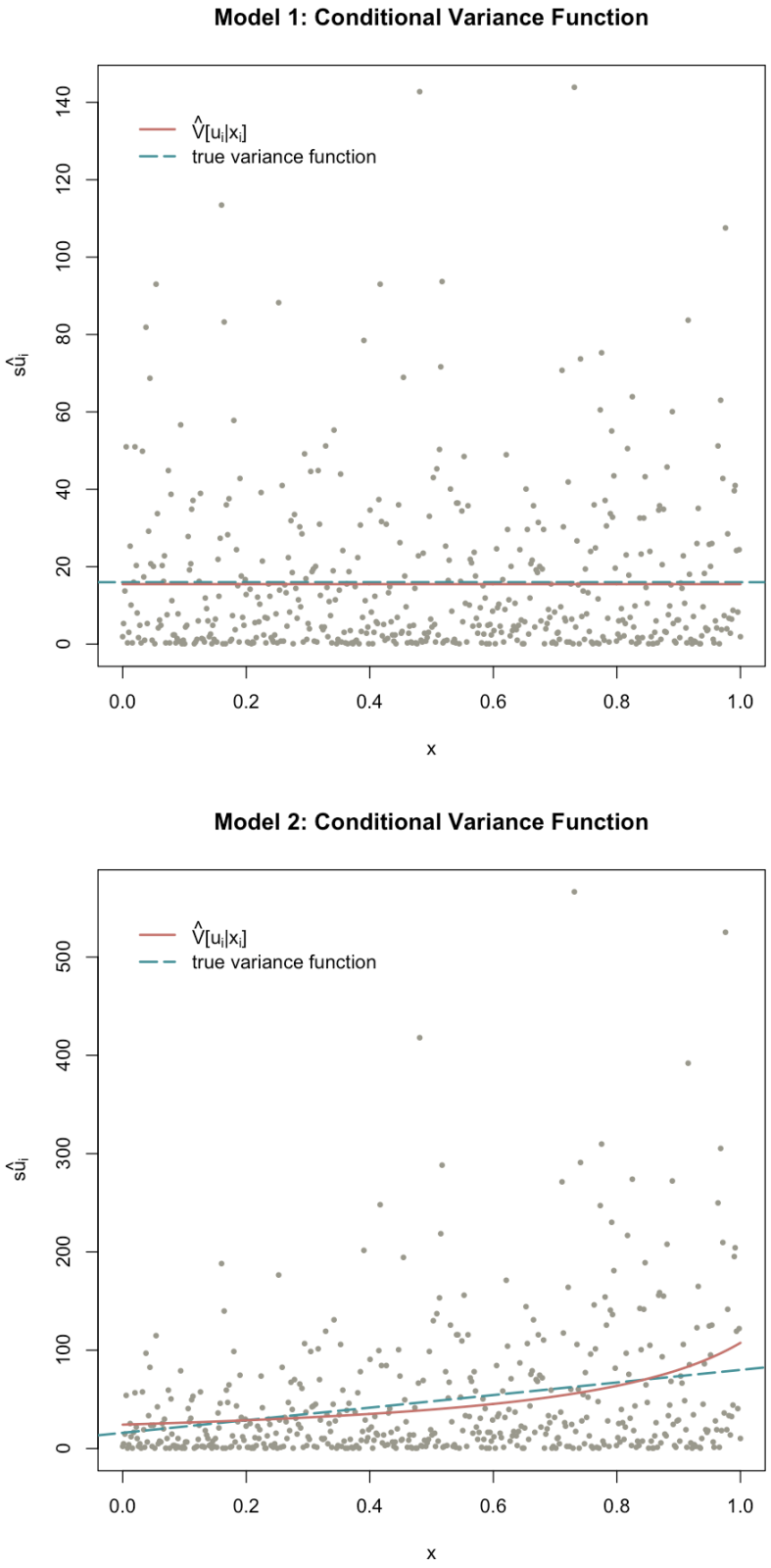


Figure 3.4: *Conditional Variance Function Estimation. The figures plot the squared centered residuals of the simulated data for the Location-Scale and Location-Scale Shift Models given in the Example 1. Solid red line on the plots represents the estimate of the conditional variance function of the residuals of* 75^{th} *quantile and the dashed blue line the true variance function.*

is constant. Assuming that $s\hat{u}_i \sim gamma$ for which $var[Y] = (E[Y])^2$, implies again, that the gamma GLM is appropriate in this situation.

Using the fitted values of the estimated conditional variance function of the residuals, we can standardise them

$$stu_i(\tau = 0.75) = \frac{u_i(\tau = 0.75)}{\sqrt{\hat{V}[u_i(\tau = 0.75)]}},$$

and by using (3.15), we can compute the kernel density estimate of the standardised residuals

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - stu_i(\tau = 0.75)}{h}\right).$$

Having density estimates of standardised residuals Figure 3.5, we can draw a sample of standardised residuals from $\hat{f}(stu)$ using the following steps:

- *Step 1*: Choose I uniformly with replacement from $\{1, \ldots, n\}$;
- *Step 2*: Generate ϵ to have probability density function K;
- *Step 3*: Set $stu_i^* = stu_I(\tau = 0.75) + h\epsilon$, or in the case when the realisations stu_i are transformed to reflect the first and second moment properties observed in the sample $\{stu_1, \ldots, stu_n\}$, *Step 3* of the algorithm should be replaced by

  $$stu_i^* = \mu_{stu_i} + (stu_i - \mu_{stu_i} + h\epsilon)/\sqrt{(1 + h^2\sigma_K^2/\sigma_{stu_i}^2)}$$

  where μ_{stu_i} and $\sigma_{stu_i}^2$ are the sample mean and variance of $\{stu_i\}$ and σ_K^2 is the variance of the kernel K.

As a following step, we construct a smooth bootstrap sample

$$y_i^* = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(\tau = 0.75) + stu_i^* \sqrt{\hat{V}[u_i(\tau = 0.75)]},$$

and re-fit the quantile regression model to the bootstrap data

$$\mathcal{Q}_{y_b^*}(\tau = 0.75|\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_b^*(\tau = 0.75), \tag{3.32}$$

to obtain a new set of parameter estimates $\hat{\boldsymbol{\beta}}_b^*(\tau = 0.75)$. This procedure is repeated B times, giving us a $B \times p$ matrix in which $\hat{\boldsymbol{\beta}}_b^*$ is a $p \times 1$ vector of quantile regression

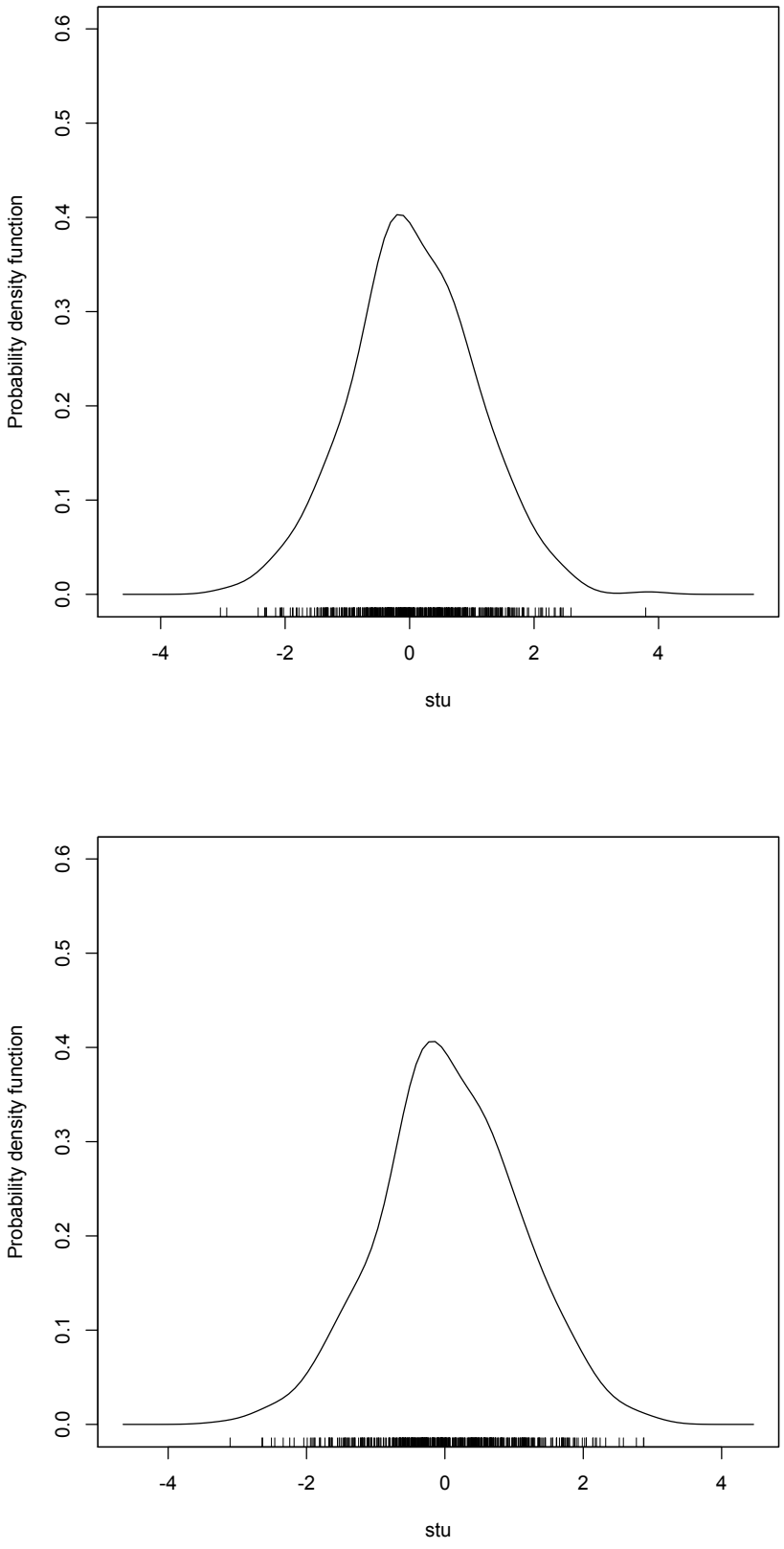


Figure 3.5: *Kernel Density Estimate. The figures plot the kernel density estimate of the standardised residuals for the two models: the location-shift and the location-scale models respectively, for the* 75^{th} *quantile regression line.*

coefficients. Using this matrix we can build up the empiricall distribution of the parameter estimates, which will enable us to make inference about $\boldsymbol{\beta}(\tau)$ (see Section 3.1).

Let us focus on the problem of the confidence intervals for the median regression ($\tau = 0.5$) parameters as it is most straight-forward when comparing the performance between the different methods and the true parameters values. We will consider the two given models: pure location-shift model (Model 1) and location-scale model (Model 2), using the following `R` based functions:

- `glm.Gamma.Inv_ks`: the conditional variance function is estimated using gamma GLM with the canonical link, after which the residuals are standardised and their density estimated using kernel smoothing.
- `glm.Gamma.Inv_ksm`: redefine kernel smooth bootstrap in which the density estimate has the same mean and the variance as the empirical data used to estimate the density.
- `xy`: $(x,\ y)$ pair bootstrap method.
- `pwy`: Parzen-Wai-Ying bootstrap method.
- `mcmb`: markov chain marginal bootstrap method.
- `wxy`: generalized bootstrap method of Bose and Chatterjee (2003) with unit exponential weights.
- `riid`: rank-score test method assuming iid errors.
- `rnid`: rank-score test method assuming nid errors.
- `wiid`: Wald method assuming iid error, with scalar sparsity estimate, using Hall and Sheather's bandwidth.
- `wker`: Wald method assuming nid error, with Powell's sandwich estimate.
- `wnid`: Wald method assuming nid error, with Siddiqui sandwich estimate, using Hall and Sheather's bandwidth.

In this simulation study, we use 500 realisations ($R = 500$) of 500 observations ($n = 500$) with $1,000$ bootstraps ($B = 1,000$). The Wald and bootstrap confidence intervals are all based on estimating standard deviations, while the rank-score intervals are computed by the parametric programming algorithm described earlier in section 2.6.

To make an informative comparison, we judge the performance of the nine listed methods by the average lengths (L) and coverage probability (C) of the computed 95% confidence intervals of the 500 realisations for each of the coefficients in the two given models.

3.4.1 Model 1: A Location-Shift Model

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Gamma.Inv_ks	2.02541	0.01879	4.96214	0.03395
glm.Gamma.Inv_ksm	2.02677	0.01875	4.96014	0.03388
xy	2.01944	0.01842	4.96510	0.03279
pwy	2.01990	0.01849	4.96422	0.03281
mcmb	2.01956	0.01848	4.96548	0.03274
wxy	2.02018	0.01847	4.96462	0.03284
riid	2.02013	0.01904	4.96211	0.03386
rnid	2.02013	0.01904	4.96211	0.03386
wiid	2.02013	0.01904	4.96211	0.03386
wker	2.02013	0.01904	4.96211	0.03386
wnid	2.02013	0.01904	4.96211	0.03386

Table 3.2: Parameter estimates for Model 1 for $\tau = 0.5$.

All the methods provide close parameter estimates, with the intercept being slightly over estimated and the slope slightly underestimated (Table 3.2). From Table 3.3 we can see that all the methods perform well. The shortest average lengths of the parameters' confidence intervals are obtained when using `riid` and `rnid` methods, with the coverage probability being the smallest compared to the other approaches used. The kernel smoothing methods provide competitive results. In particular the

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Gamma.Inv_ks	96.80%	1.83341	0.00501	96.20%	3.17149	0.00867
glm.Gamma.Inv_ksm	96.00%	1.75092	0.00469	95.20%	3.03332	0.00818
xy	96.40%	1.79039	0.01397	95.40%	3.13494	0.02213
pwy	96.60%	1.79458	0.01414	96.00%	3.13874	0.02229
mcmb	95.60%	1.77899	0.01209	94.60%	3.10114	0.02173
wxy	96.20%	1.78469	0.01408	95.00%	3.12381	0.02221
riid	89.00%	1.43243	0.01246	89.00%	2.48450	0.02108
rnid	89.00%	1.43368	0.01246	89.00%	2.49089	0.02103
wiid	95.00%	1.76606	0.01156	94.60%	3.05737	0.02002
wker	97.60%	2.05399	0.00493	98.00%	3.55384	0.00816
wnid	95.80%	1.77898	0.00859	95.80%	3.08959	0.01163

Table 3.3: Results for Model 1 for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

kernel smoothing method adjusted for the first and the second moment appear to outperform the others as the coverage and the mean lengths for both of the parameters, β_0 and β_1, are the best.

3.4.2 Model 2: A Location-Scale Shift Model

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Gamma.Inv_ks	2.03666	0.02564	4.93910	0.05749
glm.Gamma.Inv_ksm	2.03913	0.02562	4.93566	0.05739
xy	2.02679	0.02452	4.94486	0.05476
pwy	2.02757	0.02460	4.94344	0.05471
mcmb	2.02895	0.02476	4.94279	0.05532
wxy	2.02791	0.02458	4.94447	0.05482
riid	2.02698	0.02580	4.93872	0.05733
rnid	2.02698	0.02580	4.93872	0.05733
wiid	2.02698	0.02580	4.93872	0.05733
wker	2.02698	0.02580	4.93872	0.05733
wnid	2.02698	0.02580	4.93872	0.05733

Table 3.4: Parameter estimates for Model 2 for $\tau = 0.5$.

The estimation of the parameters is very close to the true values for all methods, with the intercept again being slightly over and the slope slightly under estimated

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Gamma.Inv_ks	98.60%	2.97461	0.00803	94.60%	5.14411	0.01391
glm.Gamma.Inv_ksm	98.40%	2.84057	0.00759	93.40%	4.92105	0.01326
xy	96.80%	2.38126	0.01853	94.40%	5.16402	0.03728
pwy	96.80%	2.38565	0.01864	94.60%	5.16527	0.03745
mcmb	98.00%	2.60154	0.01722	95.00%	5.23220	0.03808
wxy	96.00%	2.37212	0.01863	94.20%	5.14685	0.03734
riid	88.80%	1.88311	0.01602	87.60%	4.01254	0.03380
rnid	89.00%	1.89676	0.01593	89.00%	4.12025	0.03450
wiid	98.60%	2.84608	0.01845	92.80%	4.92709	0.03195
wker	98.40%	2.97323	0.00655	97.40%	5.92936	0.01489
wnid	96.00%	2.39943	0.01231	95.20%	5.16897	0.01935

Table 3.5: Results for Model 2 for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

(Tables 3.4). Table 3.5 reports the result from Model 2, and we can notice that in terms of coverage kernel smoothing methods works very well and that they are competitive when compared with the other bootstrapping methods. Observing the mean length of confidence intervals, we can see that for the intercept parameter, β_0, they appear to be slightly wider than those of other bootstrap methods, but quite similar to the results of the Wald method with Powell's sandwich estimate. For the slope parameter, β_1, when compared with the other bootstrap methods, kernel smoothing method is again outperforming all of the existing methods.

3.5 Conclusion

The coefficient of variation of the gamma distribution is constant, hence a good estimation of the common variance function in the homoscedastic models is expected, especially when the error is normal, as in this case: $e \sim \mathcal{N}(0,\ 16)$. Consequently, the standardisation of the residuals would work well, and as we can see from this simulation study, when compared with other existing bootstrap methods, kernel smoothing bootstrap would provide very competitive results. We have also applied the approach of pooling the residuals together to estimate their density in order to obtain smoothed

bootstrap sample. The results were the same as those obtained by estimating the density of the standardised residuals, in which they are scaled back to their original scale by using the estimated conditional variance function.

For the heteroscedastic model (M2), the variability in estimation of the intercept appears to be slightly bigger when using kernel smoothing bootstrap, than those of the other competing methods, providing slightly wider mean length for the intercept. This could be caused by an overestimated intercept of the conditional variance function, pulled by the extreme observations in the sample. Considering that we have a conditional distribution of the error, one way of investigating this issue would be to compare the bootstrap results when sampling from kernel smooth density with those when sampling from the $\mathcal{N}(0,\ 1)$.

From this simulation study Tables: 3.2, 3.3, 3.4 and 3.5, suggest that sampling from the adjusted kernel smoothing bootstrap would provide better results than the kernel smoothing bootstrap which assumes the first and the second moment of the density estimate. It shows that the kernel smoothing bootstrap method performs well when the error is normal, which automatically poses the question how robust is this method when applied to the models in which the error is not normally distributed.

Chapter 4

Extending Resampling Schemes to Non-Normal Data

4.1 Introduction

In the previous chapter we have introduced a new approach to bootstrapping using the conditional variance function for making inferences for quantile regression models. In this method we model the conditional variance function using the squared residuals from the fitted quantile function. By modeling the conditional mean of the squared residuals using gamma GLM, we are able to estimate the conditional variance function of the τ^{th} quantile under the assumption of error being normally distributed. It is of our interest to explore the robustness of this method when the error is not normal and therefore the residuals are non-gamma. Furthermore, in practice we are also not always confident that the constant coefficient of variation (CV) assumption of the gamma distribution is true, i.e.

$$CV = \frac{\sqrt{\frac{\nu}{\lambda^2}}}{\frac{\nu}{\lambda}} = \frac{1}{\sqrt{\nu}} = \phi^2.$$

Therefore, ideally, we need to extend this method to allow the change of the shape (ν) of the gamma distribution and hence enable the CV not to be constant.

4.2 Extending Gamma GLMs

The parameter $\boldsymbol{\beta}$ of the GLM can be estimated using maximum likelihood estimation. The log-likelihood of a single observation is

$$\ell(\theta_i, \phi; y_i) = \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} \right) + c(y_i, \phi)$$

where $a_i(\phi) = \phi / w_i$, and therefore, for the set of independent observations, the log-likelihood will be $\sum_i^n \ell(\theta_i, \phi; y_i)$, which can be maximised to obtain $\hat{\boldsymbol{\beta}}$.

McCullagh and Nelder (1989, Chapter 2.5) show that by applying Newton-Raphson method with Fisher scoring the optimisation is equivalent to iteratively reweighted least squares (IRWLS). Nelder and Wedderburn (1972) first proposed Fisher scoring as a general method for estimating $\boldsymbol{\beta}$ in GLMs. An initial estimate $\boldsymbol{\beta}$ is obtained and updated to $\boldsymbol{\beta}^*$ by

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \left(E\left[- \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right] \right)^{-1} \frac{\partial \ell}{\partial \boldsymbol{\beta}},$$

where both derivatives are evaluated at $\boldsymbol{\beta}$ and the expectation is evaluated using $\boldsymbol{\beta}$ as the true parameter values. After each estimation $\boldsymbol{\beta}$ is replaced by $\boldsymbol{\beta}^*$ and this procedure is repeated until convergence is obtained. The exact form for those equations is

$$\boldsymbol{\beta}^* = \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z},$$

in which $\boldsymbol{z}$ is the n-vector with i^{th} component

$$z_i = (y_i - \mu_i) g'(\mu_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

and $\boldsymbol{W}$ the $n \times n$ diagonal matrix with

$$\boldsymbol{W}_{ii} = w_i \left(g'(\mu_i)^2 b''(\theta_i) \right)^{-1}$$

Green and Silverman in their book (1994, Chapter 5.2.2), further show how the parameter ϕ cancels out during the iterative estimation of $\boldsymbol{\beta}$. This implies that fitting procedure uses only $\eta = g(\mu)$ and $V(\mu)$ and does not require knowledge of the distribution of y. This point will be shown to be of importance when estimating

the conditional variance function based on the squared residuals that are not gamma distributed.

We can summarise the above procedure by simply realising that we are regressing $g(y)$ on $\boldsymbol{X}$ with weights inversely proportional to $V\big[g(y)\big]$. This means that we linearise $g(y)$ using one step expansion:

$$\begin{aligned} g(y) &\approx g(\mu) + (y-\mu)g'(\mu) \\ &= \eta + (y-\mu)\frac{d\eta}{d\mu} \\ &\equiv z \end{aligned}$$

and

$$\hat{V}(z) = \Big(\frac{d\eta}{d\mu}\Big)^2 V(y) = \frac{1}{w}.$$

Faraway (2006, Chapter 6.2) breaks the IRWLS procedure into the following steps:

- *Step 1*: Set the initial estimates $\hat{\eta}_0$ and $\hat{\mu}_0$;
- *Step 2*: Form the "adjusted" dependent variable $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{d\eta}{d\mu}|\hat{\eta}_0$;
- *Step 3*: $w_0^{-1} = (\frac{d\eta}{d\mu})^2|\hat{\eta}_0 V(y)$;
- *Step 4*: Reestimate $\boldsymbol{\beta}$ to get $\hat{\eta}_1$;
- *Step 5*: Iterate *Steps 2-3-4* until convergence.

This means, that for gamma response we have:

$$\begin{aligned} \eta &= -\frac{1}{\mu}, \\ \frac{d\eta}{d\mu} &= \frac{1}{\mu^2}, \\ V(y) &= \frac{\mu^2}{\nu}, \\ w &= \Big[\frac{1}{\mu^2\nu}\Big]^{-1} = \mu^2\nu. \end{aligned}$$

4.3 Inference in Gamma GLMs

We assess the goodness of fit by how well the fitted values produced by the model mach the values of the data. The inadequacy can be observed by the discrepancy

between the model and the data. In GLM this measure is called *deviance* (3.28) and was introduced in the previous chapter 3.3.1. Deviance, D, is defined as 2 times the difference in log-likelihood between the full model, $\ell(y, \phi|y)$, and the model under consideration $\ell(\hat{\mu}, \phi|y)$. We have seen that the scaled deviance can be written as

$$D^*(y, \hat{\mu}) = \frac{1}{\phi} \sum_{i=1}^{n} 2w_i\big(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\tilde{\theta}_i)\big).$$

Considering that the deviance measures the closeness of the fitted values by the model to the data, it is often interpreted as the residual sum of squares in a linear model. However, Pearson's chi-squared statistic, χ^2

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{\big(y_i - E[y_i]\big)^2}{Var[y_i]} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{b''(\theta_i)} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

could be used as an alternative measure of discrepancy. Both, the scaled deviance, $D(y, \hat{\mu})/\phi$, and the Pearson's chi-squared statistic, χ^2, are asymptotically distributed as χ^2_{n-p}, where p is the number of fitted parameters in the model under consideration. McCullagh and Nelder (1989, Chapter 2.3.1 and Chapter 4.4.3) discuss the deviance function and Pearson's statistic in more details. They point out that there are some strong assumption to be made in order to expect that each has approximately χ^2_{n-p} distribution, especially after scaling with dispersion parameter ϕ.

Suppose that $\boldsymbol{Y}$ has a gamma distribution with independent components, McCullagh and Nelder (1989, Chapter 8.3.6) point out that $D(y, \hat{\mu})$ is very sensitive to rounding errors in very small observations and it is infinite if any component of $\boldsymbol{Y}$ is zero. Moreover, if the gamma assumption is false, $\hat{\phi} = \frac{1}{\hat{\nu}}$ does not consistently estimate the coefficient of variation and for that reason they suggest the moment estimator

$$\tilde{\sigma}^2 = \sum_{i=1}^{n} \frac{\big(\frac{y_i - \hat{\mu}_i}{\hat{\mu}}\big)^2}{n-p} = \frac{\chi^2}{n-p},$$

which is consistant for σ^2, assuming that $\boldsymbol{\beta}$ has been consistently estimated.

4.4 Adjusting Gamma GLMs for Non-Constant Coefficient of Variation

In this part of the study our task is to examine how effective the proposed bootstrap method is when used for the quantile regression models with non-normal errors. To estimate the conditional variance function we fit a gamma GLM to the squared residuals and in the case when the residuals are non-normal we do not know how robust the gamma GLM would be. Thus, assumption that allows us to model the conditional variance function using GLM is violated as it does not accommodate a non-constant coefficient of variation.

In classical normal linear regression modeling we often apply a transformation to the data which aims to achieve approximate normality and constant variance. We have seen that GLMs provide flexibility to this requirement as they provide a choice of distributions and enable additive systematic effect of a covariate to hold on an independently chosen transformed scale ($\eta = g(\mu)$). By making this choices we are accommodating a mean-variance relationship that is appropriate for the data. Once the variance is specified as a function of the mean, the variance is assumed known up to a constant of proportionality.

In GLMs the deviance is defined as

$$D(y, \hat{\mu}) = \phi D^*(y, \hat{\mu}) = 2\phi\Big[\ell(y, \phi) - \ell(\hat{\mu}, \phi)\Big],$$

and for $y_i \sim gamma(\nu, \lambda)$ it is

$$\phi D(y, \hat{\mu}) = 2\Big[\sum_{i=1}^{n}\Big(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\Big(\frac{y_i}{\hat{\mu}_i}\Big)\Big)\Big]. \tag{4.1}$$

Thus, having fitted a model we are able to estimate the individual deviances

$$\hat{d}_i^* = 2\left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\Big(\frac{y_i}{\hat{\mu}_i}\Big)\right], \tag{4.2}$$

which by regressing $\hat{d}_i^*$'s on x enables the estimation of the conditional dispersion function $\big(\hat{f}(\phi|\boldsymbol{x})\big)$

$$\tilde{d}_i^* = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i \tag{4.3}$$

assuming $e_i \sim \mathcal{N}(0, \sigma^2)$.

Adopting this procedure in which we are estimating the conditional dispersion function ($\hat{\phi}_i$), we can estimate the change of the shape parameter of the gamma distribution, which will allow us to adjust for the change in CV.

For gamma distribution we have

$$\theta = -\frac{1}{\mu}, \qquad \mu = \frac{\nu}{\lambda} \qquad \text{and} \qquad \phi = \frac{1}{\nu},$$

therefore the fitted values can be rescaled by the new, individual shape parameters,

$$\hat{\theta} = \frac{1}{\hat{\mu}} = \frac{\hat{\lambda}}{\hat{\nu}} \hat{\nu} \frac{1}{\hat{\nu}_i}$$

that would provide a better reflection of the true variance function.

4.4.1 Application

Let us look at a heteroscedastic, location-scale shift model used in Section 3.4 in the Example 1:

$$\text{Model 2} \quad : \quad y_i = 2 + 5x_i + \sigma(x_i)e_i$$

with $e_i \sim t(5)$ and $\sigma(x) = \sqrt{1 + 4x}$, for which we wish to estimate median quantile, $\tau = 0.5$. After fitting the required quantile and squaring the centered residuals we fit a gamma GLM in `R`

$$\texttt{V(u|x)=glm(}su_i \sim \texttt{1 + x, family=Gamma(link="inverse"))},$$

to obtain the estimate of the global shape parameter $\hat{\nu} = 0.465$ and the estimates of the individual deviance parameters $\hat{d}_i^*$ using (4.2).

Our main objective is to rescale the estimated conditional variance function to allow for change in CV, that would allow us to get a closest possible estimate of the true variance function $\sigma(x)$.

We are going to examine three different approaches of regressing $\hat{d}_i^*$ on x to see which one would provide us with the most flexibility:

i *Linear least square regression* by applying (4.3);

ii *Robust regression.* When the error distribution is not normal, linear-least square estimates can behave inadequately. In such cases we tend to use robust regression. The commonly used method of robust regression is *M-estimation*, introduced by Huber (1981). It is a form of weighted least squares regression, which uses iterative minimization of the sum of the squared residuals. The new set of weights are determined at each iteration by the estimated residuals, in the way that the larger the residuals, the smaller the weights. The weights depend on the residuals and, at the same time, the residuals depend on the model that depends on the weights. This generates an iterative process and it goes on until the change in the parameter estimates are below a preset threshold. At the end, instead of all points being weighted equally, the weights vary and those with the largest weights contribute more to the fit. For more details about the implementation of this method see Section 8.2;

iii *Non-parametric regression* using local polynomial regression (*lowess* - locally weighted scatterplot smoother, for the simple-regression) (Cleveland, 1979). We fit the model

$$\hat{d}_i^* = f(x_i) + \epsilon_i,$$

where ϵ_i are random variables with mean 0 and a constant scale, and we focus on evaluating the regression function at a particular x_0-value. We apply a p^{th}-order weighted-least-squares polynomial regression of $\hat{d}_i^*$ on x,

$$\hat{d}_i^* = b_0 + b_1(x_i - x_0) + b_2(x_i - x_0)^2 + \cdots + b_p(x_i - x_0)^p + \epsilon_i$$

weighting the observations in relation to their proximity to the focal value x_0. The larger the order of the local regressions p, the more flexible the smooth function is going to be. As weight function *the tricube function* is used

$$W(z) = \begin{cases} (1 - |z|^3)^3 & \text{for } |z| < 1 \\ 0 & \text{for } |z| \geqslant 1, \end{cases}$$

where

$$z_i = \frac{x_i - x_0}{h},$$

where h is the width of a window fencing the observations for the local polynomial regression. The fitted value at x_0 is simply $\hat{d}_i^* = b_0$. Commonly, the value of h can be adjusted in such a way that each local polynomial regression includes a fixed proportion s of the data, in which case we refer to s as *the span* of the local-regression smoother. The size of the span controls the level of smoothing: the larger the span, the smoother the result (Fox, 2002).

Figure 4.1 shows the scatter plots of the squared centered residuals of the location-scale shift model (M2) given in Example 1 (see Section 3.4), on which the true and estimated conditional variance functions are superimposed. In the first plot, we can see that the estimated variance function using gamma GLM, solid blue line, is very close to the true variance function (solid gray line) and it appears that those two lines are almost overlapping. However, when we focus more closely on this area (lower plot in Figure 4.1) we notice that our gamma GLM estimate is pulled down and is lieing below the true function, although the intercept of the line appears to be overestimated. In terms of the adjusted variance function estimates by applying linear least square regression the line is pulled upwards, so that the estimate is closer to the true line. The other two, parametric and non-parametric approaches, are pulling the estimate further down from the true line, subsequently getting the intercept closer to the true value. The true reflection of how successful those three adjustments are, we are going to be able to see after using them to standardise the residuals of the fitted quantile function and ultimately by observing the results of the bootstrap procedure in which those standardised residuals are used.

We also note that our estimated line rather than being a straight line, as expected according to the true variance function, is curved as a consequence of the canonical, inverse link used. Obtaining a straight line approximation would be possible to achieve by applying the 'identity' link in gamma GLM. However, we have tried to do so, but the gamma GLM with identity link proved not to be stable, as the fitting of such a model was failing on some of the occasions.

Another approach when using GLM, that would allow us to get an estimate that

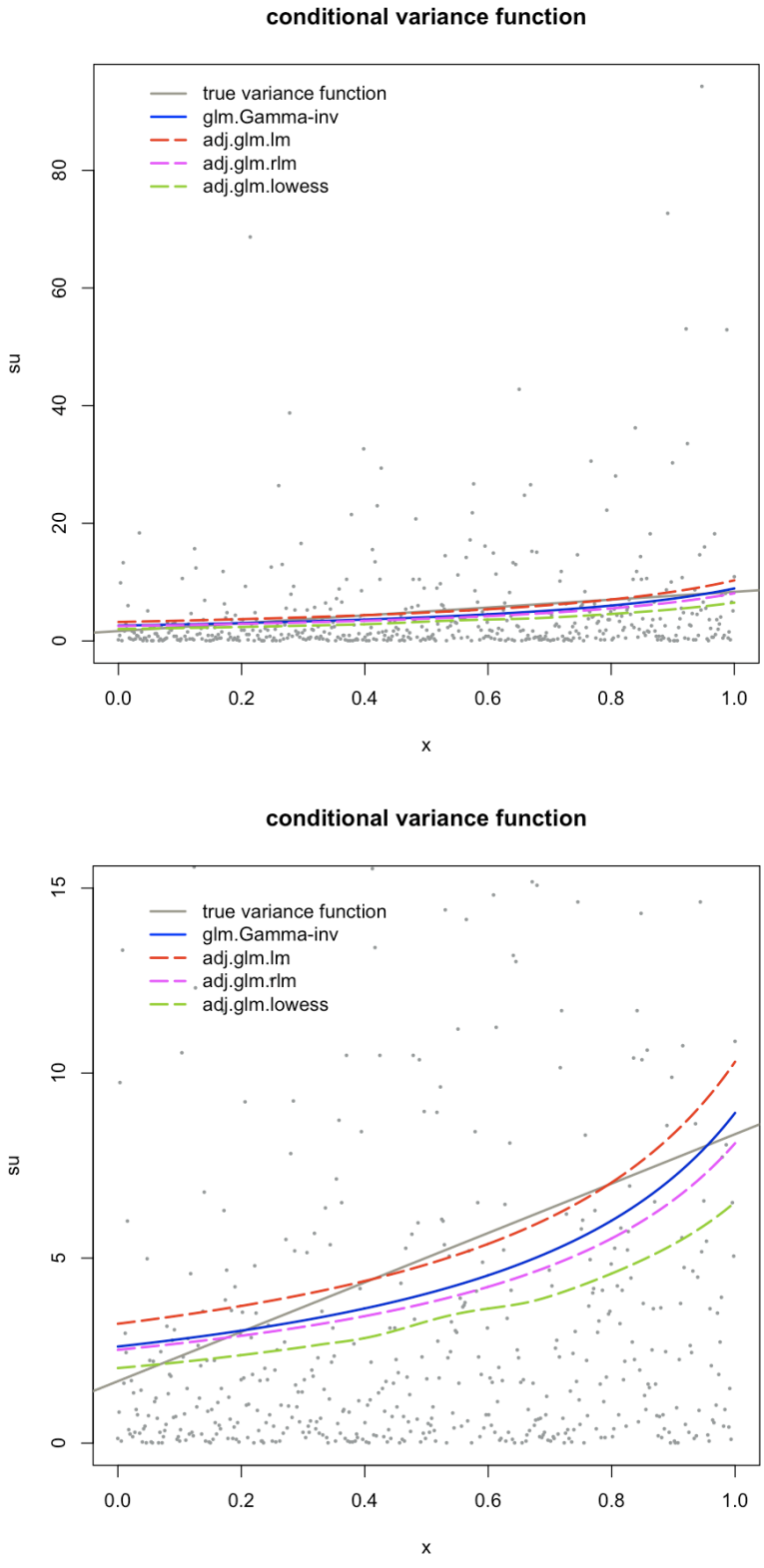


Figure 4.1: *Conditional Variance Function Estimation. The figures plot the squared centered residuals of the simulated data for the location-scale shift model (M2) given in Example 1 (Section 3.4). Solid gray line on the plot represents the true variance function* $\sigma(x) * Var(t(5))$. *The solid blue line is the estimated conditional variance function using gamma GLM model, whiles red, magenta and green dashed lines present adjusted gamma GLM estimate in which* $\hat{d}_i^*$ *is regressed on* x_i *using least-square, robust and lowess approach respectively. The lower plot shows the fitted lines in more detail.*

enables dispersion parameter ϕ_i to vary with the covariate x and that would make possible to have a straight line as an estimate of the conditional variance function by joint modelling of mean and dispersion proposed by Nelder and Lee (1991).

4.5 Joint Modelling of Mean and Dispersion

In GLM models the variance of the response is assumed to take the form

$$Var[Y_i] = \phi V(\mu_i),$$

where $V[\mu]$ is a known variance function and its choice directs the interpretation of the dispersion parameter ϕ. In the case of gamma GLM, we know that $V[\mu] = \mu^2$ and that ϕ is the squared coefficient of variation of the response. We have realised that when applying gamma GLM to estimate the conditional variance function of the non-normal error quantile regression models, we need to relax this assumption and enable ϕ_i to varies with the covariate x. Indeed, it would be more convenient to assume that ϕ_i is proportional to some known weights $w_i = 1/m_i$, in the cases where Y_i is the average of the m_i elementary observations (Nelder and Wedderburn, 1972).

Nelder and Lee (1991) proposed a model that consists of two interlinked GLMs, one for the mean and one for the dispersion. They suggest to use the deviance component from the model for the mean as the response for the dispersion model. At the same time, the fitted values of the dispersion model provide the weights for the model of the mean. To start this iterative procedure we first model the mean using $\phi = 1$ and usually five cycles suffice.

4.5.1 Joint Model Specification

The joint model is specified as a standard GLM in terms of the dependence on covariates of the first two moments. For the mean we use:

$$E[Y_i] = \mu_i \qquad \nu_i = g(\mu_i) = \sum_j x_{ij}\beta_j \qquad Var[Y_i] = \phi_i V(\mu_i) \qquad w_i = \frac{1}{\phi_i}, \qquad (4.4)$$

assuming that the observations are independent. The dispersion ϕ_i is no longer considered to be fixed and is assumed to vary. Using gamma GLM we can model

the estimate d_i of the dispersion obtained after fitting the model of the man in the following way:

$$E[d_i] = \phi_i \qquad \zeta_i = \log(\phi_i) = \sum_j z_{ij}\gamma_j \qquad Var[d_i] = \tau\phi_i^2. \tag{4.5}$$

The two models (4.4) and (4.5) are interlinked, the model for the mean will produce the response for the model for the dispersion, which on the other hand will produce the weights for the mean model.

In our study, as the response in the dispersion we use

$$d_i = \frac{(y_i - \hat{y}_i)^2}{(1 - h_i)}$$

with a gamma GLM with a log link and weights of $1 - h_i$ as suggested by Smyth et al. (2001), where h_i's are diagonal elements of the hat matrix H and y_i's are our squared residuals (su_i) of the fitted quantile function. The following code shows the implementation of the joint modelling in R:

```
> lmod<-lm(su~1+x)
> for (i in 1:5) {
lhat<-influence(lmod)$hat
d<-((residuals(lmod))^2)/(1-lhat)
mv<-glm(d~1+x, family=Gamma(link=log), weights=1-lhat)
w<-1/fitted(mv)
lmod<-lm(su~1+x, weights=w)
}
```

4.6 Example 2: Applying Smooth Bootstrap to Non-Normal Data

Let us again consider the two models given in Section 3.4 : a homoscedastic, location-shift Model 1 (M1) and a heteroscedastic, location-scale shift Model 2 (M2)

$$\text{Model 1} \quad : \quad y_i = 2 + 5x_i + e_i \quad \text{and}$$

$$\text{Model 2} \quad : \quad y_i = 2 + 5x_i + \sigma(x_i)e_i$$

used in Example 1 (Section 3.4). This time, let us have $\{e_i\}$ which is iid, but from tree different distributions:

- $\mathcal{N}(0, 16)$,
- $t(20)$ and
- $t(10)$,

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$, and where $\sigma(x) = \sqrt{1 + 4x}$. We wish to estimate 50^{th} quantile, $\tau = 0.5$, thus 50^{th} quantile of the u_i variable is at 0. For each of the six simulation studies, we will use five hundred realisations ($R = 500$) with a thousand bootstraps ($B = 1,000$). As in the previous study, to make an informative comparison, we judge the performance of the used methods by the average lengths (L) and coverage probability (C) of the computed 95% confidence intervals (CI) of the 500 realisations for each of the coefficients in the two given models.

We again focus on the median regression ($\tau = 0.5$) expanding the set of method from the previous study (see Section 3.4) by the newly introduced one:

- *IRWLS*
- *Joint Modelling*
- *Adjusted for Dispersion* using: *lm*, *rlm*, *lowess* and
- *glmGamma link-log.*

We bootstrap using kernel smoothing (*ks*) and redefine kernel smoothing (*ksm*) as before. In order to investigate the increased variability in estimation of the intercept when using kernel smoothing bootstrap, considering that we have a conditional distribution of the error, we will also resample from the $\mathcal{N}(0,\ 1)$ and $\mathcal{N}(0,\ \sigma^2(stu))$, when the error is $e \sim \mathcal{N}(0, 16)$. In the cases when the error is $e \sim t(20)$ and $e \sim t(10)$ we will resample from $t(20)$ and $t(10)$ respectively. In practice we would not know these distributions but in the simulations they provide a good comparison for the

bootstrap procedures based on kernel density estimation. Note that stu_i are standardised residulas used in the bootstrapping procedure as introduced before. The full list of the results are given in the Appendix B.

Figures 4.2 illustrates fitted variance functions to the models given in the Example 2, using seven different methods: gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log), in which error is from the three different distributions: $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$.

Before we go into the analysis of the obtained results given in the Appendix B we need to report a problem we had in some parts of this simulation study. As we have moved from the normal error and in particular from the scenario in which the errors are independent and identically distributed (iid), some of the methods we wanted to use to estimate the conditional variance function were breaking down. For the nid case where the error was from the t distribution with $df = 20$, glm.Gamma with link inverse had failed. As we moved further away from the normal distribution and wanted to use t distribution with $df = 10$ Joint Modelling mated also failed this time even in the iid case.

Figure 4.3 shows the plot of the squared residuals, su_i, from a fitted quantile regression model with the nid error from the t distribution with $df = 20$ (Model 2) and the index plot of Cook's distance for the linear model in which su_i's are regressed on x_i. We have noticed that when we fit an ordinary least squares model in which we regress su_i on x_i we can identify the influential observations which are potential creators of algorithm failure when fitting a $glm.Gamma(link = "inverse")$. Calculating the Cook's distance D_i is a commonly used way of identifying the influential observations. Cook's distance is a scale-invariant measure of the distance between the regression coefficients with the i^{th} observation absent and present. It is a measure that incorporates the measure of outlying-ness and the measure of the leverage. As a cutoff value, the value of $4/(n-k-1)$ has been suggested, where n is the number of the observation and $k+1$ the number of the regression coefficients use in the model (Fox, 2002).

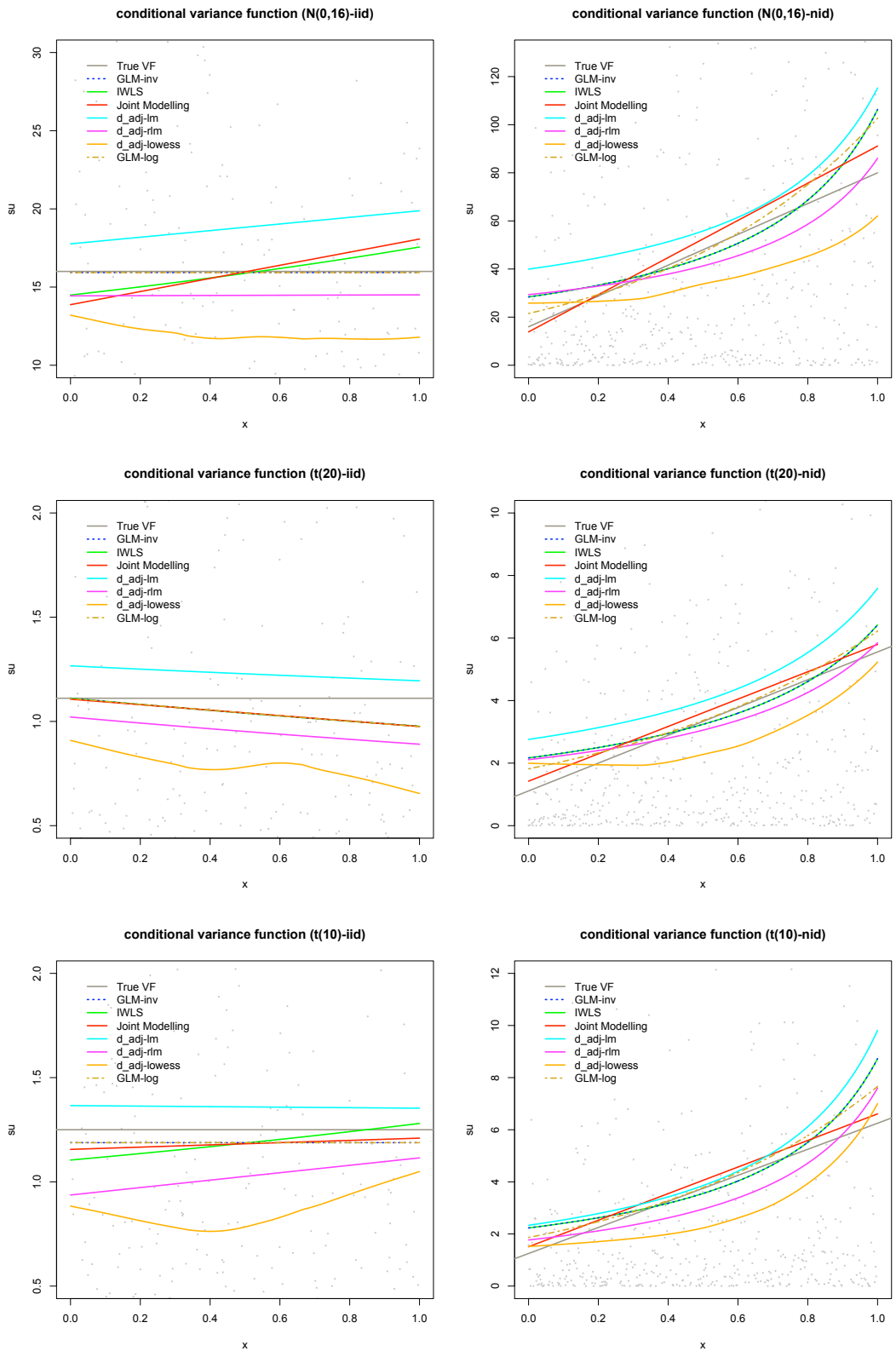


Figure 4.2: *Conditional Variance Function Estimation. The figures plot the squared centered residuals of the simulated data for the two models given in the Example (M1 and M2). Solid gray lines on the plots represents the true variance function. The estimates of the conditional variance function are given in dashed blue, green, red, turquoise, magenta, orange and dashed golden when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) respectively. The error in the two models is from three different distributions:* $\mathcal{N}(0, 16)$, $t(20)$ *and* $t(10)$.

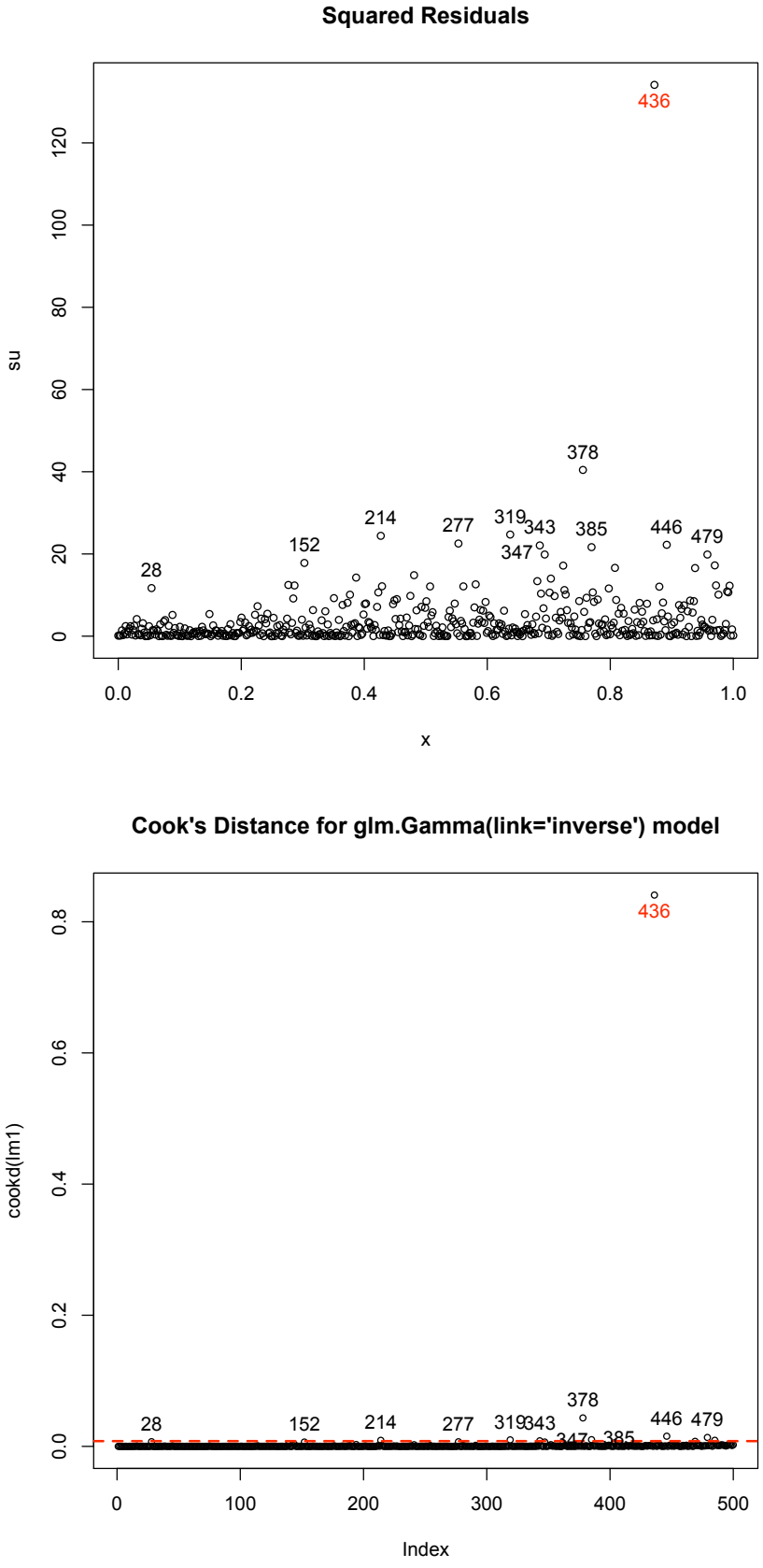


Figure 4.3: *The Outliers. First graph plots the squared residuals from the fitted quantile regression model with nid error from the t distribution with df = 20. Second graph is index plot of Cook's distance for ordinary least squares model in which squared residuals are regressed on x. The dashed, red, horizontal line presents the cutoff value of 4/(n-k-1). The observations with the heigh D_i are identified.*

Even though it appears to be relatively easy to identify the outliers using this approach it does not provide a satisfactory solution to our problem. First, it is hard to make direct connection between ordinary least squares model and a general linear model with gamma distribution, to see how the one would enable direct identification of the outliers for the other model. Even if we would use this approach far too many observations would be identified as outliers, since we have found out that for the data set example illustrated in Figure 4.3 after removing only the most extreme observation (observation no 436) we could fit a $glm.Gamma(link = "inverse")$ without the failure. We are also dealing with a heteroscedastic problem in which the variance is not constant as assumed in the calculation of the Cook's distance. Most importantly, the final flaw in identifying the outliers and removing them from the data would cause the break down in our algorithm of standardisation of the residuals as the estimates of the conditional variance function for some of the observations identified as the outliers would be missing.

In Section 8.2 we have proposed a way of identifying the outliers using robust regression in which the model is fitted by iterated re-weighted least squares (IWLS) regression using Huber's M-estimator. After fitting a cubic line to the squared residuals su_i, the weights produced in a robust regression procedure are used to identify the most extreme values. We notice that if we would remove the observations with 0 weight ($w_i = 0$), which are deemed to be extreme, we can fit $glm.Gamma(link = "inverse")$ to the remaining data without failure. Considering that for our study we need to have a full data set, data sets with observation with ($w_i = 0$) are then dismissed and another was simulated. Only the data set in which all of the observations had ($w_i \neq 0$) are included in the further simulation study.

4.6.1 Model 1: Homoscedastic, Location-Shift Model

All of the methods used for the estimation of the conditional variance function provide very good results. Estimates of the parameters are very close to the true values. In the case of the normally distributed error, the intercept parameter, b_0, is slightly

over-estimated and the slope parameter b_1 slightly under-estimated. When the error is from the t distribution with $df = 20$ the situation is opposite: b_0 is slightly under-estimated and b_1 slightly over-estimated, for all of the methods used in the study. For the error from the $t(10)$ methods for estimating the conditional variance function based on $glm.Gamma(link = "inverse")$ and $JointModelling$ were failing and we had to exclude 26.04% of the data sets from the study for which we had identified the extreme outliers (the observations with $(w_i = 0)$) as the potential causes of the methods failure (see Table 4.1). When using our approach the intercept is slightly

No of outliers	5	4	3	2	1	0
No of data sets	1	1	3	32	139	500

Table 4.1: Number of data sets with the identified extreme outliers for the data with *iid* error from $t(10)$.

over-estimated and under-esetimated in other cases, whilst the slope parameter is under-estimated for all of the methods used in the study. In all of the three models, the coverage probability for the parameters when using kernel smooth bootstrapping are very competitive in comparison to the other existing bootstrapping methods.

When applying our approaches the average length of the confidence intervals for the intercept and the slope are narrower than those of other bootstrapping methods. This is particularly the case when the density from which we are bootstrapping is adjusted for the first and the second moment of the sample used in the bootstrapping procedure. We are not surprised to see that when bootstrapping from the distribution of the error used in the model, in this case $t(10)$, we get good results, but nevertheless not as good as when bootstrapping from the kernel density estimate that has the first and the second moment properties the same as the sample used. Our methods also show another good result, which is comparatively very small standard error of the average length of the confidence intervals, that is in some cases more than twice smaller than those of the other bootstrapping methods. We also notice that the Wald-Test based approaches provide very competitive results. The Rank Score method, although provides very narrow average length of the parameters' confidence intervals,

has the smallest coverage probabilities.

4.6.2 Model 2: Heteroscedastic, Location-Scale Shift Model

In this part of the study all of the methods used provide very close parameter estimates to their true values. Kernel smooth bootstrapping method with sampling from a density with the same first and the second moment as the data used, provides the best results in terms of the narrowest confidence interval of the slope parameter, b_1, and very small variability as the standard error is less than half the size of the existing competing bootstrapping methods. However, the mean length of the intercept parameter estimate, b_0, is again slightly wider that those of the competing methods, but still has smaller standard error. As we moved further from the normally distributed error, the average length of the CI for b_0 widens slightly. We get the best results for b_0 when bootstrapping from the same distribution of the error term in the model, but not so good for b_1.

The number of data sets that was not included in the study because of identified extreme outliers with $w_i = 0$, as potential causes of the methods failure was 21.14% for the error from $t(20)$ (Table 4.2) and 55.71% for the error form $t(10)$ (Table 4.3).

No of outliers	3	2	1	0
No of data sets	1	18	127	500

Table 4.2: Number of data sets with the identified extreme outliers for the data with *nid* error from $t(20)$.

No of outliers	5	4	3	2	1	0
No of data sets	1	8	40	168	412	500

Table 4.3: Number of data sets with the identified extreme outliers for the data with *nid* error from $t(10)$.

4.7 Conclusion

As we move away from the normally distributed data the coefficient of variation (CV) obtained through the gamma GLM does not change fast enough to adequately model the changes in variance, as the variance is increasing a lot faster than the mean. In this study we have attempted to scale the CV to allow for this faster change in variance than the one predicted by the gamma GLM. We have modeled the conditional deviance function using three different approaches: *lm*, *rlm* and *lowess*, which after being used to rescale the CV of the gamma GLM had provided us, rather disappointingly, with the results that are no better than those obtained by the gamma GLM itself. If we look at the tables of mean variance of the standardised residuals for all six models (see Tables: B.3, B.6, B.9, B.12, B.15, B.18) the poorest standardisation was when applying this approach, although still close to the mean value of one. When *rlm* was used to model the dispersion over x, we were getting better results than when using *lm* and *lowess*. The *lm* approach pulled the conditional variance function line upwards and *lowess* pulled it too far down (see Figure 4.2). Thus, it appears that the estimate of the conditional variance function in our resampling approach can benefit from deflation before further bootstrapping.

For all of the homoscedastic models kernel smoothing bootstrapping has provided very good results, but when dealing with the heteroscedastic data we still have slightly over-inflated confidence interval for b_0. Of all the approaches introduced to estimate the conditional variance function, the best results are provided when using $JointModelling$, in which the residuals of the fitted linear model are rescaled by the elements of the hat matrix, $\boldsymbol{H}$, by using $(1 - h_i)$. Hence it would be useful to explore the possibilities of fitting generalized linear models by using robust methods.

One way of overcoming the problem of inadequate CV modeling by gamma GLM for non-normal error models would be to explore the modeling possibilities offered by the *Quasi-Likelihood Functions*. When using GLM the characteristic of the response distribution is reflected through the form of dependence of the variance on the mean. For quasi-likelihood GLM the precise response distribution is not specified, but rather

a link function and the form of the variance function as it depends on the mean. For some non-gamma response distribution, this would enable us to specify a link and variance function combination that does not correspond to any of the standard GLMs.

This study has shown that GLM procedures are not robust. The GLM fitting procedures were breaking down for some of the data sets with the extreme outliers. To overcome this difficulty we adopted a rather conservative approach of removing the samples with identified outliers, which is not satisfactory. The kernel soothing bootstrapping when adjusted to have the same mean and the variance as the data from which it is constructed, provided very competitive result even when the data is not normally distributed. The best results were provided when using:

- the kernel smoothing bootstrapping method adjusted to have the same mean and the variance as the data from which it is constructed (ksm) and
- the bootstrapping from the error distribution used in the model $\{(\mathcal{N}(0,1), t(20)$ and $t(10)\}$, as this is a reference, as in practice we do not know the true error distribution.

Consequently, in the following studies we recommend focusing only on those two bootstrapping methods.

It appears that $glm.Gamma$ is less sensitive to the distribution of the response than expected, as for the data samples with the error distributions with long tails, for which the outliers were not identified, the overall results obtained for the kernel smoothing bootstrapping were considered good for the reasons pointed out earlier in this chapter. However, for distributions with long tails, the gamma GLM fitting procedure was failing. Subsequently we should explore some robust GLM model fitting procedures that accommodate the modeling of the response variable having an approximate gamma distribution.

Chapter 5

Extending General Linear Models for Resampling Schemes

5.1 Introduction

By using an approach that is analogous to linear method for normal data, GLMs allow us to model responses which are non-normally distributed. What makes GLMs more general is that they can model a mean-variance relationship appropriate to the data and in that an appropriate scale can be chosen for modeling the mean as a linear function of the covariates. The down side of the GLMs, that poses a trouble to us, is that once we specify the relationship between the mean and the variance, the variance is assumed known up to a constant of proportionality, to which we refer as dispersion parameter. What we need is a more flexible GLM that would allow for non-homogeneous dispersion.

In the attempt to address this problem we are going to explore the options provided by *Double Generalised Linear Models*, which model the mean and dispersion simultaneously in the context of a general linear model, allowing for heteroscedastic dispersion (Smyth and Verbyla, 1999a). To avoid the break down in the model fitting procedure when assuming that the squared residuals, su_i's, of a fitted quantile regression faction model, $\mathcal{Q}(\tau)$, are gamma, we will consider using robust estimation.

We are also going to relax the assumption that the su_i's are strictly gamma

distributed by assuming that they are from a *Tweedie* family of distributions (Smyth and Jorgensen, 1999). Finally, we will step aside from gamma GLMs all together and investigate the flexibilities offered by the *Quasi-Likelihood* GLMs.

5.2 Double Generalised Linear Model

Double Generalised Linear Model (DGLM), introduced by Smyth (1989), differ from a classical GLM in the ability to model the dependance of the dispersion parameter on the covariates in the same way as does the mean. The parameters are estimated using maximum likelihood when the population is normal, inverse Gaussian or gamma and by using quasi-likelihoods it could be adopted for the other distributions.

Let heave independent responses y_i, $i = 1, \ldots, n$ and covariates $\boldsymbol{x}$ and $\boldsymbol{z}$ with unequal weights w_i. The GLMS expect the density of y_i to be of the following form

$$f(y; \mu, \phi/w_i) = c(y, \phi/w_i) \exp\left[\frac{w_i}{\phi}(y\theta_i - b(\theta_i))\right] \tag{5.1}$$

for some specific functions b and c (McCullagh and Nelder, 1989; Section 2). In Section 3.3.1 we have seen that $\mu_i = E(y_i) = b'(\theta_i)$ and $Var(y_i) = (\phi/w_i)V(\mu)$, where $V(\mu) = b''(\theta_i)$ is a variance function of the mean-variance relationship. Parameter ϕ measures the variability in y_i once the weights and the mean-variance relationship has been taken into account.

DGLMs commence with the premise that each observation may have its own link function

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{5.2}$$

where g is a monotonic, differentiable function, $\boldsymbol{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of unknown parameters (see Section 3.3.1). It further assumes that the variance can be also expressed in the same manner

$$Var[y_i] = \phi_i w_i^{-1} V(\mu_i), \tag{5.3}$$

in which V is a standard, positive, variance function determined by the probability distribution, the w_i are known weights so that the despersion ϕ_i can be modeled as

$$h(\phi_i) = \boldsymbol{z}_i^\top \boldsymbol{\lambda}. \tag{5.4}$$

In (5.4) h is analogue to g, representing a link function, $\boldsymbol{z_i}$ is a vector of covariates affecting the dispersion and $\boldsymbol{\lambda}$ a vector of unknown parameters.

Smyth and Verbyla (1999b) suggest a more informative way of writing the distribution of y_i given by equation (5.1) as

$$f(y;\mu,\phi) = k(y,\phi)\exp\Big[-\frac{1}{2\phi}d(y,\mu)\Big], \tag{5.5}$$

in which d measures a distance between y and μ. We have seen that for the distributions covered by GLMs, d can be obtained as

$$d(y_i,\mu_i) = 2w_i\Big[t(y_i,y_i) - t(y_i,\mu_i)\Big], \tag{5.6}$$

where $t(y,\mu) = y\theta_i - b(\theta)$ bringing us to the elements we have seen in the equation (3.30): $\big\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\big\}/\phi$. For the fixed μ_i we can used weighted unit deviances $d_i = w_i d(y_i,\mu_i)$ and obtain the likelihood ratio statistics as

$$\Delta = \sum_{i=1}^{n}\phi^{-1}d(y_i,\mu_i). \tag{5.7}$$

The weighted unit deviances, d_i's, can be used as the response to estimate ϕ. This suggests that $f(y;\mu,\phi)$, given in (5.5), is an exponential density family and that

$$k(y,\phi) = \Big(2\pi\phi Var[y]\Big)^{-1/2}\rho(\phi) \tag{5.8}$$

for some function $\rho(\phi)$. The normal, inverse-Gaussian and gamma are the only exponential family dispersion models with this exact form. For the normal and inverse-Gaussian families $\rho(\phi) = 1$ and for the gamma

$$\rho(\phi) = (2\pi)^{1/2}\frac{(-\alpha)^{-(\alpha+1/2)}\exp(\alpha)}{\Gamma(-\alpha)}, \tag{5.9}$$

where $\alpha = -1/\phi$. Density given by (5.5), could be adopted for the other distributions by saddle-point approximation for exponential dispersion models, implying that

$$k(y,\phi) = \Big(2\pi\phi Var[y]\Big)^{-1/2}\big\{1 + O(\phi)\big\}. \tag{5.10}$$

This suggests that $\rho(\phi) \to 1$ when $\phi \to 0$. This saddle-point approximation (5.10) ensures that for $\phi \to 0$, $d_i \sim \phi_i\chi^2_1$. Smyth and Verbila (1999c) demonstrate how this can be seen from the following moment generating function

$$\begin{aligned} M(T) &= E\Big[\exp(T\frac{w}{\phi}d(y,\mu)\Big] \\ &= \int \exp\left(T\frac{w}{\phi}d(y,\mu)\right)\left(\frac{w}{2\pi\phi Var[y]}\right)^{1/2}\exp\left(-\frac{w}{2\phi}d(y,\mu)\right)dy \\ &= \big(1-2T\big)^{-1/2}\int\left(\frac{(1-2T)w}{2\pi\phi Var[y]}\right)^{1/2}\exp\left(-\frac{(1-2T)w}{2\phi}d(y,\mu)\right)dy \\ &= \big(1-2T\big)^{-1/2}. \end{aligned}$$

This implies that d_i follow approximately gamma GLM with mean ϕ_i, $E[d_i] \approx \phi_i$ and variance function $Var[d_i] \approx 2\phi_i^2$, link function $h(\phi)$, linear predictor $\boldsymbol{z}_i^\top \boldsymbol{\lambda}$ and the dispersion parameter 2.

Since the χ_1^2 distribution is a special case of gamma distribution, the parameters of the two submodels, $\boldsymbol{\beta}$ from (5.2) and $\boldsymbol{\lambda}$ from (5.4), can be estimated simultaneously, as the mean and dispersion of a GLM are orthogonal. For some working values for $\boldsymbol{\lambda}$, we estimate $\boldsymbol{\beta}$ using standard GLM for y_i with weights w_i/ϕ_i, to which we refer to as *the mean submodel* and vice versa. For any working value for $\boldsymbol{\beta}$, parameters $\boldsymbol{\lambda}$ are estimated using gamma GLM in which the response is d_i, which we call *the dispersion submodel*.

5.3 Tweedie Familly

Linear exponential families with a dispersion parameter (5.1), are also known as Exponential Dispersion Models (EDMs) (Dunn and Smyth, 2001). EDMs with power mean-variance relationship, $V(y) = \phi\mu^p$, where $p \in (-\infty, 0] \bigcup [1, \infty)$ are those that belong to *the Tweedie familie* (for more details see Jorgensen, 1987). For Y from a Tweedie distribution with mean μ, dispersion ϕ and variance function $V(\mu) = \mu^p$ we would write $Y \sim ED_p(\mu, \phi)$. The Tweedie family includes discrete and continuous densities, but also mixed densities as well.

The most eminent distributions of the Tweedie family are those when $1 < p < 2$,

for which the distributions are continuous for $Y > 0$ and have discrete mass at $Y = 0$

$$P(Y = 0) = \exp\Big(- \frac{\mu^{2-p}}{\phi(2-p)} \Big). \tag{5.11}$$

Jorgensen calls them *the Compound Poisson*, whilst Smyth refers to them as *the Poisson-gamma distributions*. For p equal to 0, 1, 2 and 3 we have normal, Poisson, gamma and inverse-Gausian distribution, which are the special cases of the Tweedie Family. Apart from those special cases, Tweedie densities are not known in closed form. Instead they have simple cumulant generating function (cdf)

$$B(t) = \frac{1}{\phi}\Big[b(\theta + \phi t) - b(\theta)\Big], \tag{5.12}$$

where $b(\theta)$ is the cumulant function, because when $\phi = 1$ the derivatives of b give the successive cumulants of the distribution, such as the mean, $\mu = b'(\theta)$ and the variance of the distribution $\phi b''(\theta)$ (Dunn and Smyth, 2001).

Dunn and Smyth (2005) point out that for Tweedie EDMs, the cumulant function $b(\theta)$ and the mean μ can be found by solving $b''(\theta) = d\mu/d\theta = \mu^p$ for μ and b, by setting arbitrary constant of integration to zero:

$$\theta = \begin{cases} \mu^{1-p}/(1-p), & p \neq 1 \\ \log \mu, & p = 1 \end{cases} \tag{5.13}$$

and

$$b(\theta) = \begin{cases} \mu^{2-p}/(2-p), & p \neq 2 \\ \log \mu, & p = 2. \end{cases} \tag{5.14}$$

This implies that when specifying a Tweedie GLM or DGLM we can choose between the canonical link, which is $1-p$, or a $\log -link$. It also means that the variance and the link functions can be chosen separately when using the Tweedie model approach, providing more flexibility when building a model for the data that do not necessarily correspond to any of the standard GLMs. However, to us of most interest is the power mean-variance relationship, in particular for $p \approx 2$ that corresponds to the gamma GLM.

Parameters ϕ and p of a Tweedie model can be estimated using maximum likelihood. For $Y_i \sim ED(\mu_i, \phi_i/w_i)$ it is possible to obtain the maximum likelihood

estimate of ϕ by using (5.7), where d_i is the unit variance. Dunn and Smyth (2005) indicate that for cases in which ED is not normal, inverse-Gaussin or gamma, the unit variance are not adequate for ϕ and the maximum likelihood estimator of ϕ must be computed iteratively from the full data and that the estimation of the power parameter p is even more difficult than estimation of $\boldsymbol{\beta}$ and ϕ. Nevertheless, they suggest that by being able to compute the maximum likelihood estimate of ϕ conditional on p, the MLE of p and the approximate confidence intervals can be obtained by evaluating the profile likelihood for p on a range of values followed by a univariate optimisation.

When dealing with our problem, one of the concerns is that in the quantile regression models with non normal errors, the squared residuals are not gamma distributed and perhaps $p = 2$ is not adequate for our data. Table 5.1 provides the maximum likelihood estimates of the Tweedie index parameter power, $\hat{p}$, for the models given in Example 2 (Section 4.6). Figure 5.1 contains the plots that illustrate the profile log-likelihood computed at each requested value of the $p \in [1.8, 2.2]$. The results show that for all 6 cases in our example the estimated power parameter is very close to the value of 2 ($\hat{p} \approx 2$), which is also covered by the estimated confidence intervals (see Figure 5.1). Considering that the parameter p is orthogonal to μ and ϕ in the Tweedie model (for more details see Dunn and Smyth, 2005), implies that the estimator of p changes relatively slowly as $\hat{\mu}$ and $\hat{\phi}$ change. This suggests that we can hold p fixed to 2 in our study in which case the Tweedie model reduces to an ordinary generalised model which could be observed as a prototype of a gamma GLM.

$\{e_i\}$	Homoscedastic	Heteroscedastic
$\mathcal{N}(0, 16)$	1.98776	1.99592
$t(20)$	2.02857	2.00408
$t(10)$	2.06122	2.03674

Table 5.1: Index Parameter Power p. The maximum likelihood estimates of the Tweedie index parameter power, $\hat{p}$, for the homoscedastic model: $y_i = 2 + 5x_i + e_i$ and heteroscedastic model: $y_i = 2 + 5x_i + \sigma(x_i)e_i$ with the e_i from $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$.

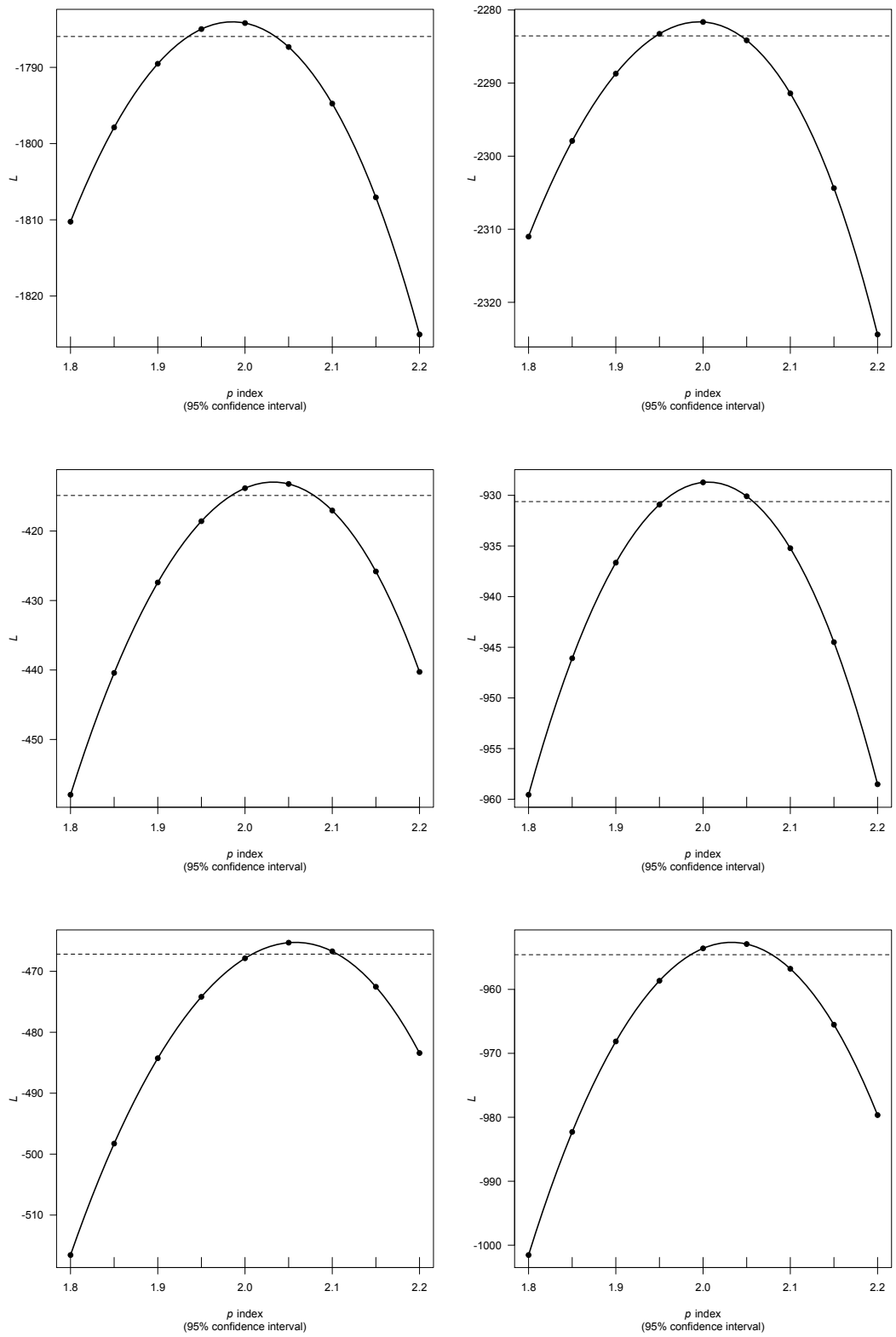


Figure 5.1: *The profile log-likelihood. The plots show the profile log-likelihood computed at each requested value for the Tweedie index parameter power p for the homoscedastic model: $y_i = 2 + 5x_i + e_i$ in column one and heteroscedastic model: $y_i = 2 + 5x_i + \sigma(x_i)e_i$ in column 2, with the e_i from $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$ in eah of the rows.*

5.4 Quasi-Likelihood General Linear Model

Quasi-likelihood generalized linear models introduced by Wedderburn (1974) have widened the scope of GLMs by allowing the full distributional assumption about the random component in the model to be replaced by the assumption in which only the first two moments are defined. This means that we can specify the link and the variance functions of the model for data with an unspecified distribution. We can define the first two moments, even though we do not have a strong idea about the appropriate distributional form for the response.

If we know the distribution of the observations, the parameters of the model can be estimated by using the method of maximum likelihood. This means that for a specified identity link and constant variance we would be able to estimate the regression parameters using least squares methods and to do the inference we would assume a Gaussian distribution of the response. The inference in this case would be fairly robust to non-normality as the sample size increases and the most important part of the model specification are the link and variance functions. Faraway (2006), points out that the same effect holds for other standard GLMS. Provided that a large sample is available the result are not sensitive to the smaller deviations from the distributional assumption. This holds the key to the idea of the quasi-likelihood GLMs, for which we need to be able only to specify the mean-variance relationship.

Let y_i $(i = 1, \ldots, n)$ be a set of independent observations with mean μ_i and variance $V(\mu_i)$, where V is a known function. Each expectation of y_i is some known function of a set of parameters $\beta_1, \ldots, \beta_2$. We can define the score function $U(y_i, \mu_i)$ as

$$U(y_i, \mu_i) = \frac{y_i - \mu_i}{V(\mu_i)} \tag{5.15}$$

which shares the following properties:

$$\begin{aligned} E(U) &= 0, \\ var(U) &= \frac{1}{V(\mu_i)}, \\ -E\Big[\frac{\partial U}{\partial \mu_i}\Big] &= \frac{1}{V(\mu_i)} \end{aligned}$$

with a log-likelihood derivative. The properties given suggest that we can define

$$Q(\mu; y) = \int_y^\mu \frac{y_i - t}{V(t)} dt, \tag{5.16}$$

as the *quasi-likelihood*, or more precisely as the log quasi-likelihood for μ based on y (McCullagh and Nelder, 1989, Capter 9). We can also define *the log quasi likelihood* for all n independent observations as the sum of the individual contributions

$$Q(\mu; y) = \sum_{i=1}^{n} Q_i(\mu_i; y_i). \tag{5.17}$$

McCullagh and Nelder (1989, Chapeter 9) provide a table with the examples of quasi-likelihood functions for a number of common variance functions, where some of them correspond to real log likelihoods for known exponential familly distributions. The standard GLM link functions are implemented as links available for quasi family in R.

The inference procedures are similar to those for standard GLM (for more details see Wedderburn, 1974; McCullagh and Nelder, 1989). By the analogy to GLM's deviance

$$D(y, \hat{\mu}) = 2\phi\Big[\ell(y, \phi) - \ell(\hat{\mu}, \phi)\Big],$$

the quasi-deviance function corresponding to a single observation is

$$D(y; y) = -2\sigma^2 Q(\mu; y) = 2\int_\mu^y \frac{y - t}{V(t)} dt,$$

because the contribution from the *full* model is zero. We can get the total deviance $D(\boldsymbol{y}; \boldsymbol{\mu})$ by the summation of the components, which is a computable function depending only on $\boldsymbol{y}$ and $\boldsymbol{\mu}$.

5.4.1 Application

Let us go back to a heteroscedastic, location-scale shift model used in in Section 3.4 in the Example 1:

$$\text{Model 2} \quad : \quad y_i = 2 + 5x_i + \sigma(x_i)e_i$$

with $e_i \sim \mathcal{N}(0, 16)$, and $\sigma(x) = \sqrt{1 + 4x}$, for which we wish to estimate median quantile, $\tau = 0.5$. After fitting the required quantile and squaring the centered residuals we fit a gamma GLM

$$\texttt{V(u|x)=glm(}su_i \sim \texttt{1 + x, family=Gamma(link="inverse"))}.$$

The residual deviance is 1256.600 on 498 degrees of freedom, with Pearson's statistic equal to 936.112. Hence the estimate of the dispersion parameter is $\hat{\phi} = 936.112/498 = 1.880$, indicating high variability in the data. Figure 5.2 shows the variability in the data, which appears not to be symmetric around the zero line and it has a higher

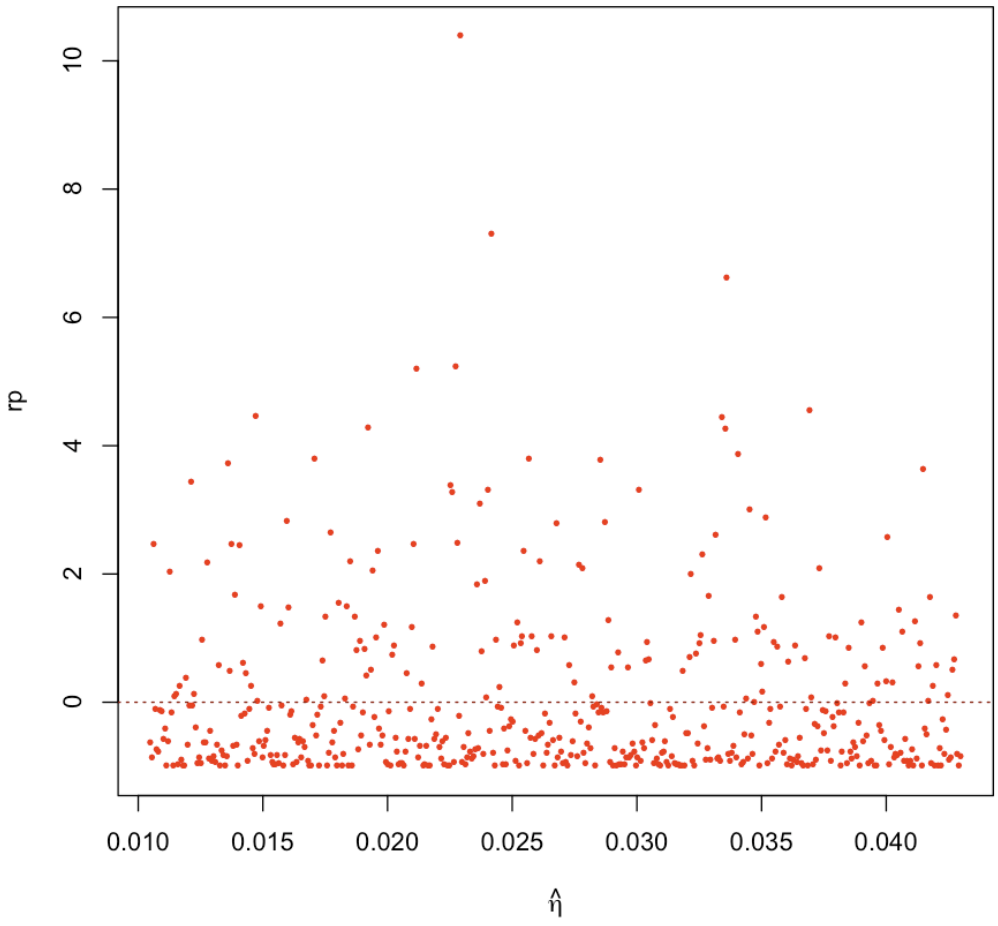


Figure 5.2: *Pearson's Residuals. For the gamma fit to the quantile's squared residual, Pearson's residuals are plotted against the linear predictor* $\hat{\eta} = 1/\mu^2$.

spread around the middle. This suggests that the chosen variance function is not a

true description of the variability in the data and a non standard GLM link function should be tried. Considering that the variability appears to be smaller at the end of the scale, we will replace standard gamma link function by a new one $\mu^2(1-\mu)^2$, in the attempt of better description of the variability of the data. The obtained quasi likelihood function of the new link function is

$$Q(\mu, y) = (2y-1)\log\left(\frac{y(1-\mu)}{\mu(1-y)}\right) + \frac{y}{\mu} + \frac{1-y}{1-\mu} - 2, \tag{5.18}$$

which we can implement in R. When fitting this quasi GLM we will use the identity link that, so far, has proven to be unstable for gamma GLM as the fitting procedure was failing. Considering that the underlying variance function in our model is linear this should provide a better fit, which Figure 5.3 illustrates.

Figure 5.4 plots the residuals of the quasi model against the linear predictor, showing slight improvement in the shape from the one shown in Figure 5.2. The residuals are still not symmetric around the zero line, but the effect of a higher spread around the middle is ostensibly smaller. However, despite the apparently better fit of the variance function which can be also seen in Figure 5.3, the estimated dispersion parameter, of the newly fitted, quasi GLM model is $\hat{\phi} = 1.951$ which is even higher than the previous one, implying the variability of the data has not been adequately described.

This leads us into a conclusion that quasi GLM would be a difficult approach to use in our simulation study as the choice of the variance function and the link function can be done only by a careful consideration based on the construction of the appropriate plots. As we have seen in our application, the improvement in terms of capturing the true variability of the data by using appropriate quasi GLM is not task that can be easily automated.

5.5 Robust General Linear Model

From the previous study, we have noticed that the gamma GLM modeling of the conditional variance function for the models with the non-normal error was not stable

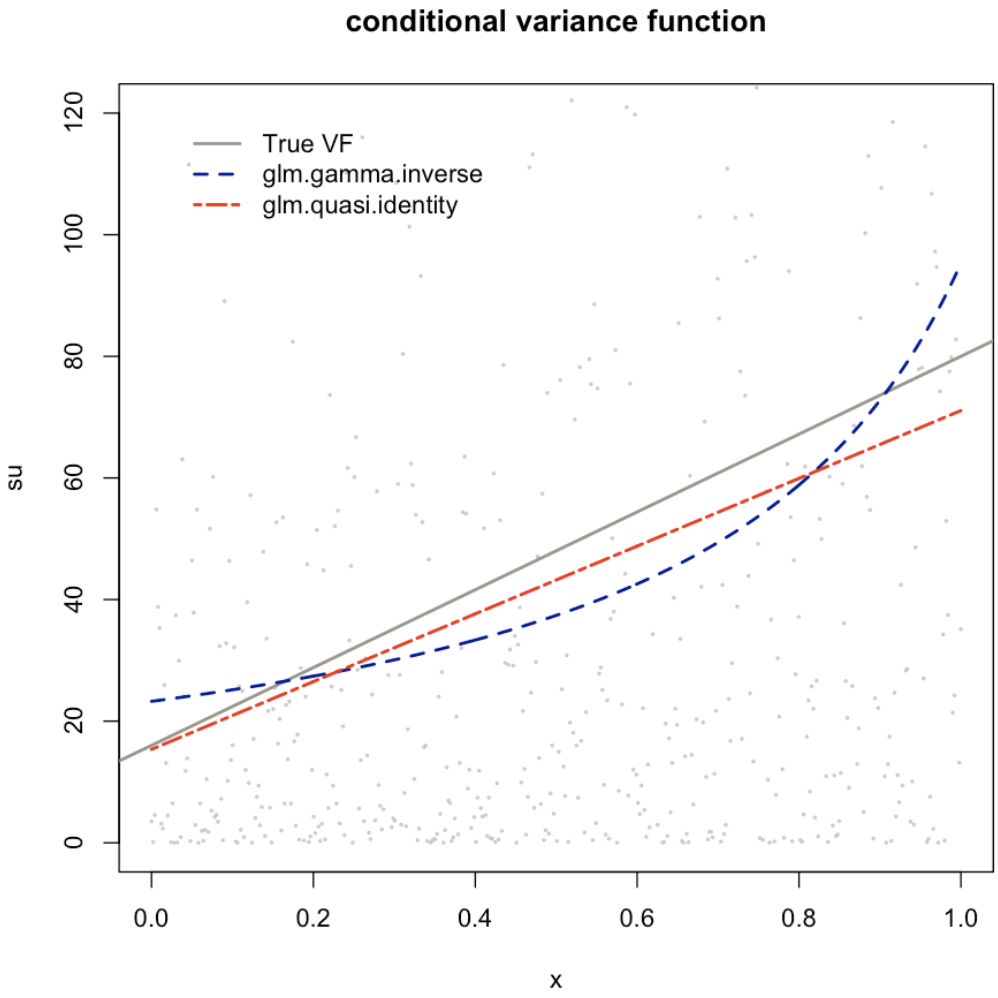


Figure 5.3: *Conditional Variance Function. The figures plots the squared centered residuals of the simulated data for the heteroscedastic model given. The solid gray line represents the true variance function, the dashed blue line and dashed red line are the estimated conditional variance functions using gamma GLM and quasi GLM respectively.*

and it was breaking down. For the error distribution with the long tails, such as $t(10)$, we have the situations where a gamma GLM is our canonical choice, but the model does not inevitably fit the data due to overdispersion. In those cases, quite too often we had identified the extreme observation that could cause the break down of the GLM model fitting procedure and those data sets were not included in the study, which was unsatisfying. What we need in those cases is a robust approach that would suppress ill effect from the outliers and enable GLM algorithm to work without failing.

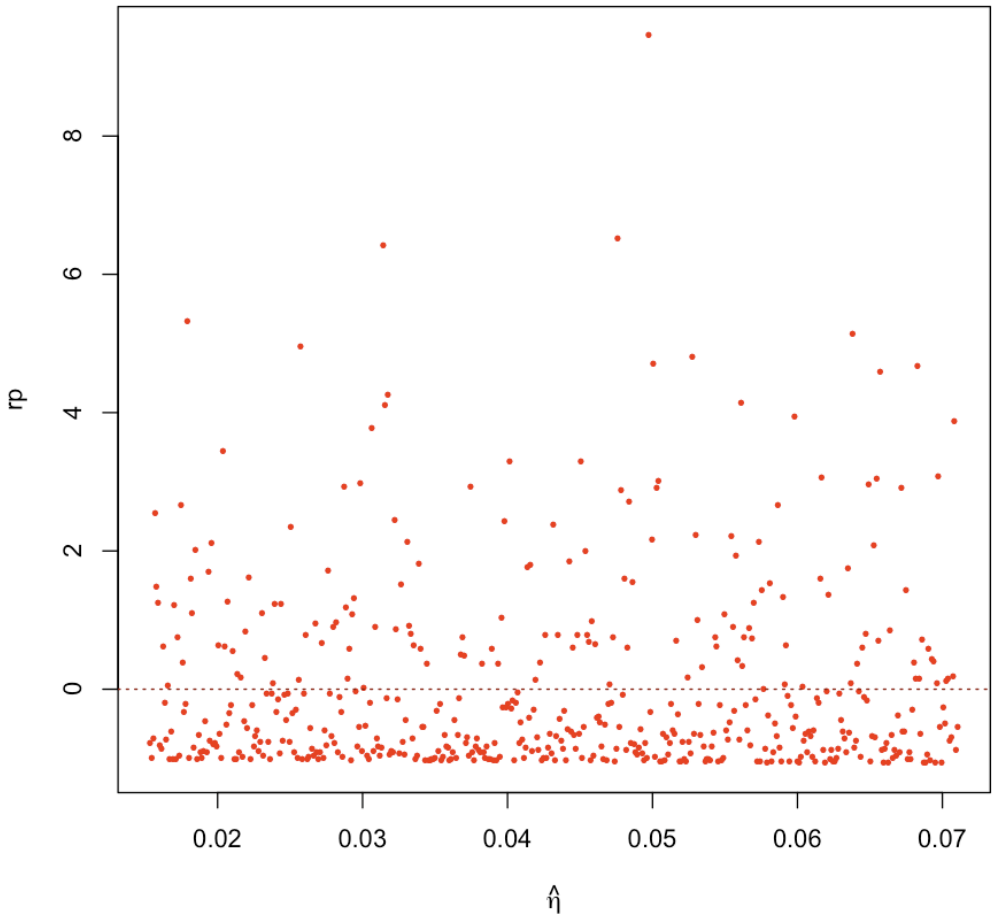


Figure 5.4: *Pearson's Residuals. For the quasi GLM fit to the quantile's squared residual, Pearson's residuals are plotted against the linear predictor* $\hat{\eta} = \mu^2(1-\mu)^2$.

Cantoni and Ronchetti (2001) have proposed a robust approach to GLM estimators based on robust deviances that are natural generalization of the quasi-likelihood functions. This quasi likelihood estimator is the solution of the system of estimating equations

$$\sum_{i=1}^{n} \frac{\partial Q(y_i, \mu_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \mu_i' = 0, \tag{5.19}$$

where

$$\mu_i' = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \tag{5.20}$$

and $Q(y_i, \mu_i)$ is a quasi-likelihood function. The solution of (5.19) is a M-estimator

introduced by Huber (1981, Section 8.2). The estimator is the solution of the estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mu_i) = \mathbf{0}, \tag{5.21}$$

where each of those equations is a M-estimator characterized by a score function

$$\boldsymbol{\psi}(y_i, \mu_i) = \nu(y_i, \mu_i) w(\boldsymbol{x}_i) \mu_i^\top - a(\boldsymbol{\beta}) \tag{5.22}$$

and where

$$a(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} E\big[\nu(y_i, \mu_i)\big] w(\boldsymbol{x}_i) \mu_i^\top. \tag{5.23}$$

The expectation in (5.23) is taken with respect to the conditional distribution $y|\boldsymbol{x}$. We can observe the estimating equation (5.21) as a special case of the link function $\eta_i = g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for GLM in which the response y_i comes from an exponential family of distributions such that $E[y_i] = \mu_i$ and $Var[y_i] = V(\mu_i)$, for $i = 1, \ldots, n$ (see Section 3.3.1). The functions ν, w are weight functions, such that when linked to the standard GLM link function (3.24) we have $\nu(y, \mu) w(\boldsymbol{x}) = V^{-1}(\mu) w(\boldsymbol{x}, y, \boldsymbol{\beta})(y - \mu)$.

The bounded function $\nu(y, \mu)$ in (5.22) controls the deviation in y space and the down-weighting of the leverage points is ensured through the weights $w(\boldsymbol{x})$. Cantoni and Ronchetti (2001) as a choice for functions ν suggest

$$\nu(y_i, \mu_i) = \psi_c(r_i) \frac{1}{V^{1/2}(\mu_i)}, \tag{5.24}$$

where

$$r_i = \frac{y_i - \mu_i}{V^{1/2}(\mu_i)} \tag{5.25}$$

are the Pearson residuals and ψ_c is a Huber function defines as

$$\psi_c(r) = \begin{cases} r, & |r| \leqslant c, \\ c\, sign(r), & |r| > c. \end{cases} \tag{5.26}$$

As a choice for the weight function w they suggest $w(\boldsymbol{x}_i) = \sqrt{1 - h_i}$, where h_i is the i^{th} diagonal element of the hat matrix $H = X(X^\top X)^{-1} X^\top$.

For $w(\boldsymbol{x}_i) = 1$, for all i and the choice of $\nu(y_i, \mu_i) = (y_i - \mu_i)/V(\mu_i)$, we would get the quasi-likelihood estimator (5.19). This implies, that for a particular choice of the functions ν and w we would actually deal with a robust equivalent of the classical quasi-likelihood function.

5.6 Example 3: Extending General Linear Models for Resampling Schemes

Lets go back to the two models given in Section 4.6 : a homoscedastic, location-shift model (M1) and a heteroscedastic, location-scale shift model (M2)

$$\begin{aligned} \text{Model 1} \quad &: \quad y_i = 2 + 5x_i + e_i \quad \text{and} \\ \text{Model 2} \quad &: \quad y_i = 2 + 5x_i + \sigma(x_i)e_i \end{aligned}$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$, and with $\{e_i\}$ which is iid from tree different distributions: $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$, and where $\sigma(x) = \sqrt{1 + 4x}$. For each of the six simulation studies, we will again use five hundred realisations ($R = 500$) of five hundred observations ($n = 500$), with a thousand bootstraps ($B = 1,000$). As in the previous studies, to make an informative comparison, we judge the performance of the used methods by the average lengths (L) and coverage probability (C) of the computed 95% confidence intervals (CI) of the 500 realisations for each of the coefficients in the two given models.

We focus on the median regression ($\tau = 0.5$) and to estimate the conditional variance function, we introduce in this simulation study the following set of methods:

- *glm.Tweedie*
- *dglm.Tweedie*
- *Adjusted glm.Tweedie for the dispersion using the dispersion model obtained by dglm.Tweedie*
- *glmrob.Gamma.*

We have explored all possible varieties within each method and the list of adopted approaches for each of the six investigated models is given in Table 5.2. Figures 5.5 illustrates fitted conditional variance functions for the six models in the Example 3, adopting the approaches as indicated in Table 5.2. The bootstrapping methods that will be used are the kernel smoothing bootstrapping adjusted to have the same mean

and the variance as the data from which it is constructed (ksm) and the bootstrapping from the corresponding error distributions: $\mathcal{N}(0,1)$, $t(20)$ and $t(10)$. The full list of the results are given in the Appendix C.

Prior to going into analysis of the obtained results for this simulation study given in the Appendix C, we should review the list of the methods used for the conditional variance function estimation. In the previous simulation study in Chapter 4.6, we had tried to address the issue of non constant CV obtained through the gamma GLM. In this study we will try to do the same, scaling CV obtained when using Tweedie GLM by Tweedie DGLM's dispersion submodel to allow for this faster change in variance than the one predicted by the GLM.

Table 5.2 reveals the robustness of the GLM and DGLM models in which the assumption of squared residuals su_i of being strictly gamma was relaxed by assuming that they are from a Tweedie family of distributions with the power $p = 2$. Acting as a prototype of gamma GLM, it appears that Tweedie GLM delivers a more robust GLM model fitting procedure.

The robust gamma GLM method works only in the cases when the error is normally distributed. As soon as we move away from the normal distribution, except for $\boldsymbol{e} \sim t(20)$ but only in the iid case, the robust method is failing as the data is too contaminated. Too many extreme values cause break down in the algorithm of the robust gamma GLM, making this approach unsuitable for the models with non-normal errors.

5.6.1 Model 1: Homoscedastic, Location-Shift Model

We have already pointed out, that when estimating the conditional variance function for the model with error from $t(10)$, robust gamma GLM method was breaking down as the data is too dispersed and too many observations could be observed as the outliers. The working methods (see Table 5.2) have provided very good results for homoscedastic models. The estimates of the parameters are again very close to the true values. In the case of the normally distributed error and the error from $t(10)$, the

No	Method	u_i					
		$\mathcal{N}(0,16)$-iid	$\mathcal{N}(0,16)$-nid	$t(20)$-iid	$t(20)$-nid	$t(10)$-iid	$t(10)$-nid
1	glm.Gamma.inverse	=	✓	=	failed	failed	failed
2	glm.Tweedie.0	= ✓	✓	= ✓	✓	= ✓	✓
3	glm.Tweedie.1	=	failed	=	failed	=	failed
4	dglm.Tweedie.0	= ✓	✓	✓	✓	✓	✓
5	dglm.Tweedie.1	=	failed	failed	failed	failed	failed
6	glm.Gamma.inverse by dglm.Tweedie.0	=	=	=	NA	NA	NA
7	glm.Gamma.inverse by dglm.Tweedie.1	=	NA	NA	NA	NA	NA
8	glm.Tweedie.0 by dglm.Tweedie.0	= ✓	= ✓	= ✓	✓	✓	✓
9	glm.Tweedie.1 by dglm.Tweedie.1	=	NA	NA	NA	NA	NA
10	glmrob.Gamma.inverse	=	✓	=	failed	failed	failed
11	glmrob.Gamma.log	= ✓	✓	= ✓	failed	failed	failed

Table 5.2: Shows what are the selected approaches for conditional variance function estimation for the models given in the Example 3. Symbol ✓ - indicates selection of the method for the given model; = - indicates that the obtained estimate is the same as the other estimates within the relevant group; **failed** - indicates that the fitting procedure algorithm was breaking for a given method and **NA** - approach was not available because the failure in the fitting procedures required.

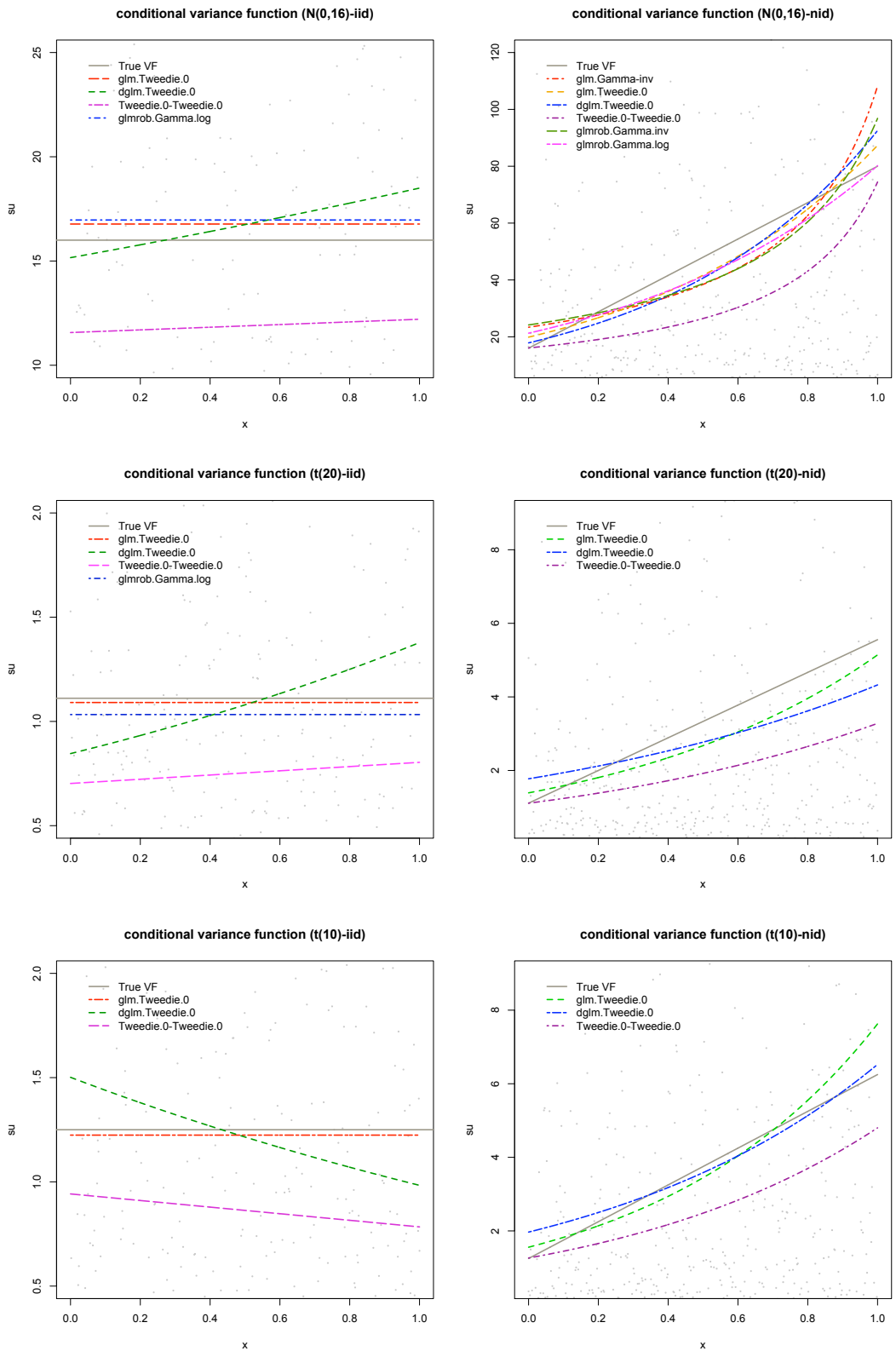


Figure 5.5: *Conditional Variance Function Estimations for Example 3 models. The figures plot the squared centered residuals of the simulated data for the two models given in the Example (M1 and M2) with the errors from the three different distributions:* $\mathcal{N}(0, 16)$, $t(20)$ *and* $t(10)$. *Solid gray line on the plots represents the true variance function. The estimates of the conditional variance functions are given in different colored dashed lines when using methods indicated in Table* 5.2 *for each of the six models.*

estimated intercept parameter, b_0, is slightly under-estimated, whilst the slope parameter, b_1, slightly over-estimated and this is the case for all of the used approaches. If we examine the results for Model 1 with $\boldsymbol{e} \sim t(20)$, we can notice the opposite effect, in which all of the used approaches have provided slightly over-estimated intercept and slightly under-estimated slope.

The coverage probabilities in the three models, for both estimated parameters, b_0 and b_1, are very competitive between all of the bootstrapping methods used. Using our approaches we get narrower mean lengths confidence intervals for the intercept and the slope parameters then when using other competing bootstrapping methods. The standard errors of the mean length confidence intervals of the parameters are at the same time much smaller in our approaches then those of the others.

Taking into consideration not just the mean lengths of the confidence intervals of the intercept and the slope parameters of Model 1, but their standard errors and the coverage probabilities for all three models, our approaches appear to perform better than any of those obtained by other competing methods. Scaling the CV obtained from Tweedie GLM by Tweedie DGLM's dispersion submodel when estimating conditional variance function did not work as well as expected. Using this approach when bootstrapping from the normal or the appropriate t distribution we get the narrowest average lengths of the confidence intervals for both parameters. However, the coverage probability is lower than in any other bootstrapping method. From Tables: C.3, C.9, C.15, in the Appendix C, we can see that this method for estimating the conditional variance gives highest mean of the variance of the standardised residuals. This tells us that out of all approaches used in this study for estimating the conditional varaince, this one will be the weakest.

5.6.2 Model 2: Heteroscedastic, Location-Scale Model

The robust gamma GLM method used in the estimation of the conditional variance function was breaking down for the models with the error from the t distribution in which we have too dispersed data. The methods used have provided very competative

results. The intercept and the slope parameters estimates are very close to their true values. In cases with $\boldsymbol{e} \sim \mathcal{N}(0, 16)$ and $\boldsymbol{e} \sim t(20)$, the intercept parameter is slightly over-estimated and the slope parameter slightly under-estimated, whilst for the model in which $\boldsymbol{e} \sim t(10)$ it is the other way around.

In all three models the coverage probabilities of the parameter estimates b_0 and b_1 are very similar for all of the approaches used, with the exception to the rank-score methods for which it is the smallest. Regardless of the bootstrapping method we use, the mean length of the confidence intervals of the intercept parameter is slightly wider when using our approaches than those of the competing methods. If we look at the probability density functions of the standardised residuals obtained from one of the simulations run in this study we can see that the standardisation works well for each of the three heteroscedastic models. The *pdf*'s of the standardised residuals for all threee models are very close to the superimposed, standard normal *pdf* on the plots for all of the approaches used (Figures 5.6 and 5.7). There are too many extreme values, introducing the bias in the model and pulling the conditional variance function up causing overestimation of the intercept parameter. Nevertheless, if we look at the standard errors of the mean lengths of the CI's, they are much smaller when using our approaches, making our estimates more consistent in length. Examining the average length of the slope's confidence intervals, its standard error and coverage probability, our approach again provide better results than other used in the study as the mean lengths are narrower with high coverage probability and much smaller variability.

When scaling the CV obtained by Tweedie GLM with Tweedie DGLM's dispersion submodel we get the highest mean variance of the standardised residuals. Nonertheless, when using this approach for conditional variance estimation and when bootstrapping from the distribution of the error term in the model we get very narrow average lengths of the confidence intervals for both parameters in all three models.

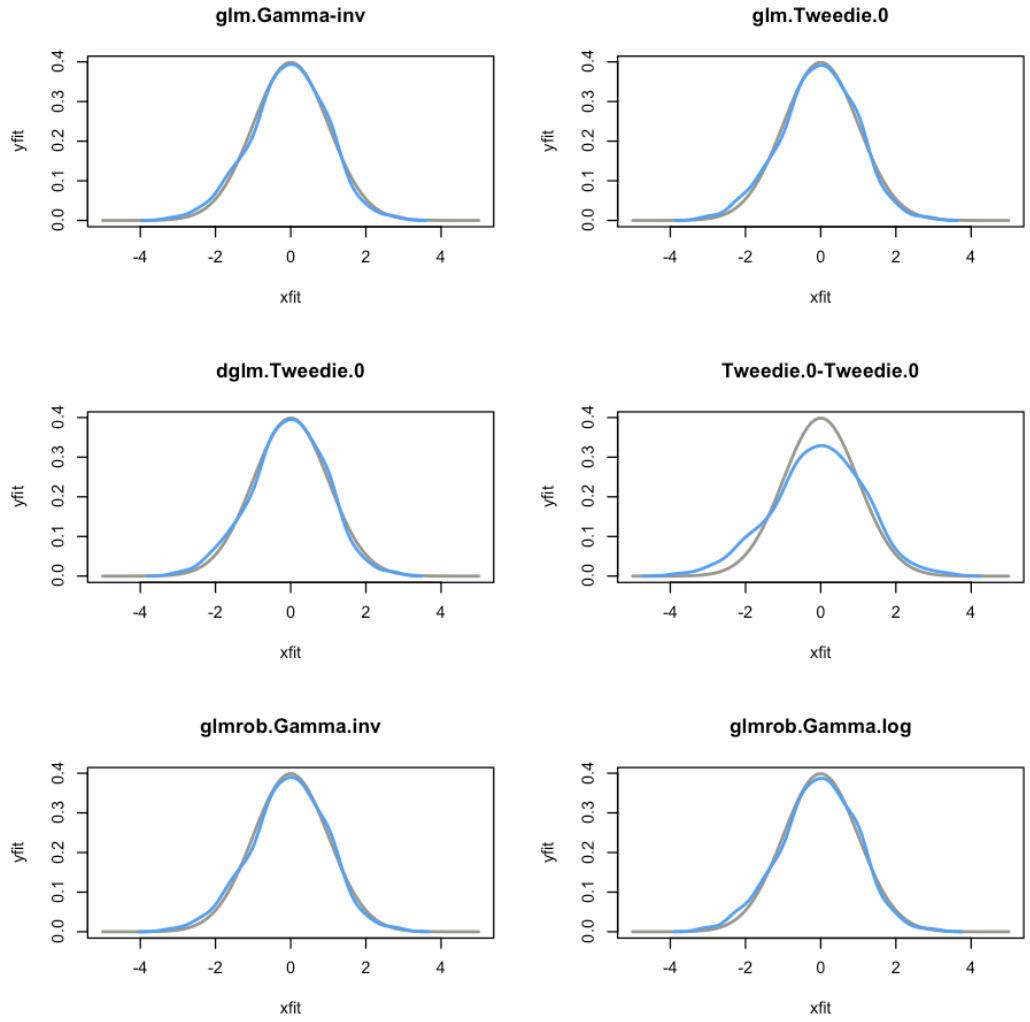


Figure 5.6: *Probability Density Functions. The pdf's of the standardised residual for the heteroscedastic model with the normal error. The residuals are standardised by the estimated conditional variance function when using glm.Gamma(link = "inverse"), glm.Tweedie(link = "log"), dglm.Tweedie(link = "log"), scaling the CV of glm.Tweedie(link = "log") by dispersion model obtained from dglm.Tweedie(link = "log"), robustglm.Gamma(link = "inverse") and robustglm.Gamma(link = "log") approaches. The pdf's of standardised residuals are plotted with a blue line and on the plots are superimposed pdf's of the standard normal distribution as gray lines.*

5.7 Conclusion

Through this simulation study we have identified, assuming that the response is from a Tweedie family of distributions with the power parameter value equal to 2, we are able to obtain a robust GLM. By using Tweedie GLM with $p = 2$ we are mimicking gamma GLM, which acts as its prototype. If we use a canonical inverse link with Tweedue GLM, we could only estimate the conditional variance function

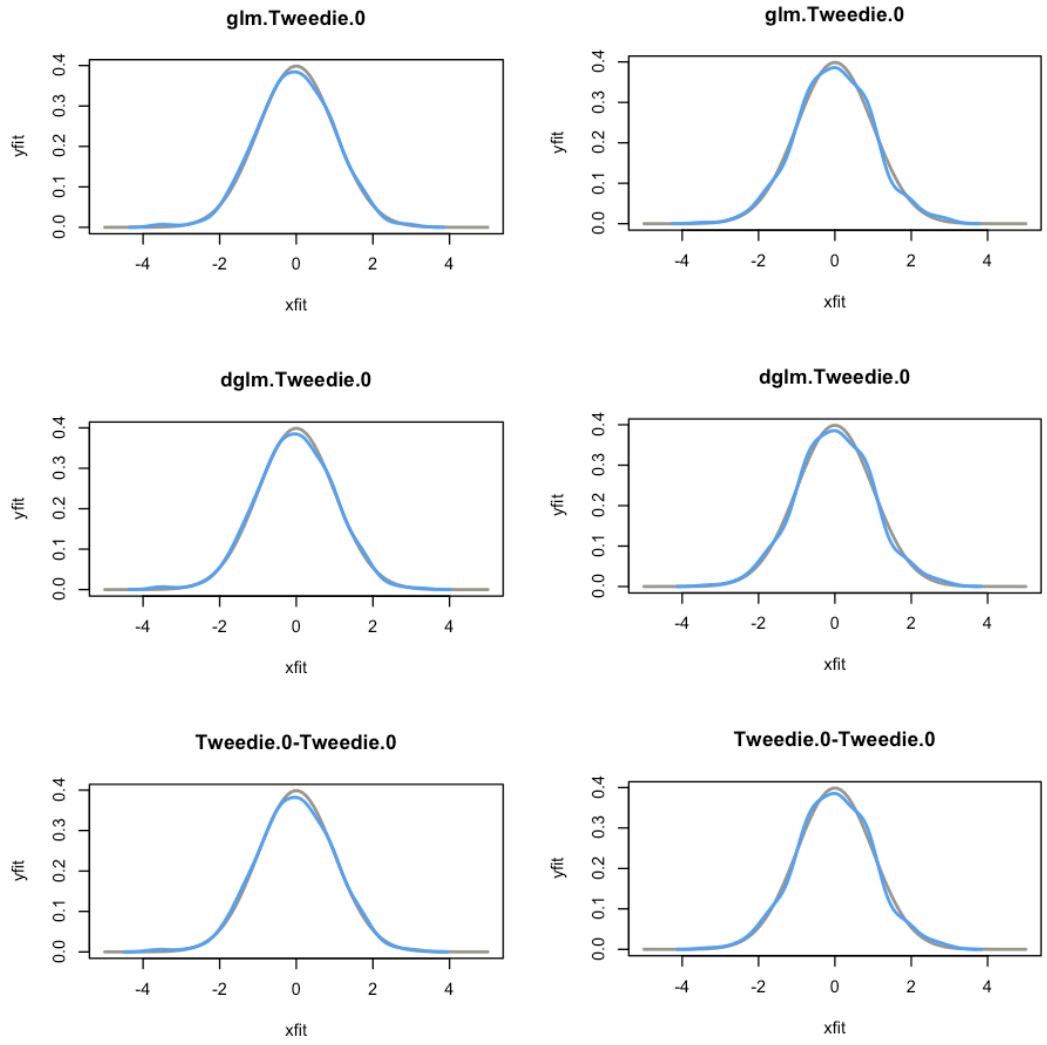


Figure 5.7: *Probability Density Functions. The pdf's of the standardised residual for the heteroscedastic models with the error from $t(20)$ and $t(10)$, given in each column respectively. The residuals are standardised by the estimated conditional variance function when using $glm.Gamma(link = "inverse")$, $glm.Tweedie(link = "log")$, $dglm.Tweedie(link = "log")$ and scaling the CV of $glm.Tweedie(link = "log")$ by dispersion model obtained from $dglm.Tweedie(link = "log")$ approaches. The pdf's of standardised residuals are plotted with a blue line and on the plots are superimposed pdf's of the standard normal distribution as gray lines.*

in the homoscedastic models. For the heteroscedastic models we had to use the *log* link. In the homoscedastic models in which both links were possible, the estimated conditional variance functions were the same.

Another method that has also proved to be robust to the distributional assumption of the response when fitting the model for the conditional variance function was DGLM with Tweedie family. This model fitting procedure broke down when using

the canonical inverse link for the models used in the simulation study, except for the homoscedastic model with normal error. The *log* link method worked without breaking down.

We still have the problem of over inflated intercept parameter estimates in the heteroscedastic models, for which we get slightly wider mean confidence interval, although the variation is significantly smaller. The approach in which we were scaling the CV obtained by Tweedie GLM with Tweedie DGLM's dispersion submodel was producing the value of the mean variance of the standardised residuals higher than 1, suggesting it is more variable. Nevertheless, this approach was pulling the estimate of the conditional variance function down, making the mean length of the confidence interval of the intercept parameter narrower when bootstrapping from the distribution of the error term of the model.

Through this simulation study we have explored parametric ways of estimating the conditional variance function, needed to standardised the residuals of the fitted quantile function prior to introducing bootstrapping procedure. We propose a way of fitting pseudo gamma GLM that is stable and robust to non-strictly gamma distributed responses by using Tweedie family of distributions. Applying kernel smoothing bootstrapping method to homoscedastic models, provides short mean lengths of the estimated parameters' CI's with small standard errors. However, assuming approximate gamma distribution of the squared residuals in heteroscedastics models causes the estimated conditional variance function to be pulled upward over-inflating the confidence interval of the estimated intercept. Using some non-parametric procedures for estimating the conditional variance function in which particular assumptions on the distribution of the response are not required may provide an alternative approach to solving this problem.

Chapter 6

Using Non-Parametric Regression Models for Resampling Schemes

6.1 Introduction

In this section we will consider non-parametric estimation of the conditional variance function used for the standardisation of the estimated quantile function residuals required by the kernel smoothing bootstrapping procedure. The approach will be based on smoothing the squared residuals by applying local polynomial fitting.

Considering residual based estimator of the conditional variance is not a new idea (Hall and Carroll, 1989; Neumann, 1994; Fan and Yao, 1998), but its implementation and application is still not fully explored. In general, as a basis of the variance estimator they use estimated residuals, which are based on the simultaneous estimation of the mean regression function by some smoothing technique. With these residuals, a kernel estimator of the variance can be obtained.

The problem we deal with does not require non-parametric estimation of the mean regression function as the residuals are already available from the fitted quantile function. However, we will explore the following step in which the available residuals will be used for the variance function estimation by applying local polynomial regression. As a more robust non-parametric regression technique, we will also investigate the use of *lowess* (used earlier for the modeling of the dispersion parameter of the gamma

GLM in Section 4.4.1), locally weighted smoothing as an outlier resistant method based on local polynomial fits.

6.2 Local Polynomial Regression

Local polynomial fitting is a generalisation of the kernel estimation. The data is split into small overlapping sections and the behavior of the mean function is modeled in each section by a polynomial of the given order p. Each data section has a neighborhood of a size h, to which we refer to as a *bandwidth*. When fitting a local polynomial function of order p, a kernel function is incorporated to give more weights to the neighboring points. By jointly putting together the estimates from each data section the overall curve is formed.

Let us denote $m(x) = E(Y|X = x)$ the conditional mean function, which we wish to estimate by using a random sample of bivariate data $(x_1, Y_1), \ldots, (x_n, Y_n)$, where the data come from a joint *pdf* $f(x, y)$. Using Taylor's series and for the purpose of easy illustration assuming that $x \in (0, 1)$, then $m(x)$ can be approximated at a given point x_0, where x is close to x_0, as

$$\begin{aligned} m(x) &\approx m(x_0) + m'(x_0)(x - x_0) + \frac{m^{(2)}(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p \\ &= \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \cdots + \beta_p(x - x_0)^p, \end{aligned} \tag{6.1}$$

provided that all the required derivatives exist. Considering that this is a polynomial of order p, (6.1), we can use this in a minimization problem that uses the data X and Y. The data is used to estimate this polynomial of order p that best estimates the unknown mean function $m(x)$ in a neighborhood of size h around point x_0. This local polynomial regression function is estimated by minimizing with respect to $\beta_0, \beta_1, \ldots, \beta_p$ the following function

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1(x_i - x_0) - \cdots - \beta_p(x_i - x_0)^p\}^2 K\Big(\frac{x_i - x_0}{h}\Big). \tag{6.2}$$

The minimizing parameters β_j, $(j = 1, \ldots, p)$, reveal the polynomial which best approximates the data in each neighborhood of a chosen size h, for the chosen order

of the polynomial p and selected kernel function K. This is done using weighted least squares methods in which the weights are assigned by the kernel function K.

To present this in the matrix notation we can define the following vectors and matrices:

$$\boldsymbol{X}_{x_0} = \begin{bmatrix} 1, & (x_1 - x_0), & \cdots & (x_1 - x_0)^p \\ 1, & (x_2 - x_0), & \cdots & (x_2 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1, & (x_n - x_0), & \cdots & (x_n - x_0)^p \end{bmatrix}$$

$$\boldsymbol{Y} = \begin{bmatrix} Y_1, & Y_2, & \ldots, & Y_n \end{bmatrix}^\top$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1, & \beta_2, & \ldots, & \beta_p \end{bmatrix}^\top$$

$$\boldsymbol{W}_{x_0} = \begin{bmatrix} K\left(\frac{x_1 - x_0}{h}\right) & 0 & \cdots & 0 \\ 0 & K\left(\frac{x_2 - x_0}{h}\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K\left(\frac{x_n - x_0}{h}\right) \end{bmatrix}$$

The least squares problem becomes to minimize the weighted sum of squares function with respect to β

$$(\boldsymbol{Y} - \boldsymbol{X}_{x_0}\boldsymbol{\beta})^\top \boldsymbol{W}_{x_0}(\boldsymbol{Y} - \boldsymbol{X}_{x_0}\boldsymbol{\beta}).$$

The solution, which minimizes the problem is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_{x_0}^\top \boldsymbol{W}_{x_0}\boldsymbol{X}_{x_0})^{-1}(\boldsymbol{X}_{x_0}^\top \boldsymbol{W}_{x_0}\boldsymbol{Y})$$

provided that $(\boldsymbol{X}_{x_0}^\top \boldsymbol{W}_{x_0}\boldsymbol{X}_{x_0})$ is a nonsingular matrix (Taylor, 2008).

In order to approximate the mean function $m(x)$ at a point x_0, we use the fitted intercept parameter, $\hat{\beta}_0$ which defines the position of the local polynomial curve at the point x_0. This means

$$\hat{m}(x_0) = \boldsymbol{e}_1^\top (\boldsymbol{X}_{x_0}^\top \boldsymbol{W}_{x_0}\boldsymbol{X}_{x_0})^{-1}(\boldsymbol{X}_{x_0}^\top \boldsymbol{W}_{x_0}\boldsymbol{Y}),$$

where the vector $\boldsymbol{e}$ is of length $p + 1$ with 1 in the first position and zeros elsewhere.

A special case is a local constant regression for the case of $p = 0$. Having $p = 0$, means that the mean function $m(x)$ is estimated locally with the constant. For $p = 0$ we have the *Nadaraya-Watson* estimator

$$\begin{aligned}\hat{m}(x_0) &= \frac{\sum_{i=1}^{n} K\big(\frac{x_i - x_0}{h}\big) Y_i}{\sum_{i=1}^{n} K\big(\frac{x_i - x_0}{h}\big)} \\ &= \frac{\sum_{i=1}^{n} K\big(\frac{x_0 - x_i}{h}\big) Y_i}{\sum_{i=1}^{n} K\big(\frac{x_0 - x_i}{h}\big)} \\ &= \sum_{i=1}^{n} \boldsymbol{W}_{h_x}(x_0, x_i) Y_i.\end{aligned}$$

We can observe Nadaraya -Watson kernel estimator as a weighted least squares estimate of the local constant β_0 (Foster, 2010).

It is well known that the performance of any kernel smoothing estimator, including local polynomial fitting, depends on the choice of the bandwidth h. We can formally define the bandwidth h as a size of the local neighborhood of the point x, which give us a window: $(x - h, x + h)$, for some small h. We can conclude that the larger the bandwidth, the more observations are covered by the window and the smoother the resulting estimate will be. Selecting too large a bandwidth would cause over smoothing by reducing the variance at the expense of creating excessive modeling bias. On the other hand, for a very small choice of h, the majority of the underlying structure of the data around the given x will be more closely followed. This would reduce the bias at the expense of increased variability, making the resulting estimate too wiggly. The dilemma we face when choosing the size of the h is that by having a narrow window around x with a very few data points would not allow to filter out the noise in the data, although it would enable the estimate to be very close to the true function $m(x)$. In order to make a balanced choice when selecting the size of a bandwidth, we have to consider the idea of bias-variance trade-off.

A common measure of discrepancy of the estimated function $\hat{m}(x)$ from the true mean function $m(x)$ at a single point is the *mean squares error*, defined by

$$MSE_x\big(m(x)\big) = E\big[\hat{m}(x) - m(x)\big]^2.$$

We can split up the mean square error into a sum of the squared bias and the variance

at a single point x

$$\begin{aligned} MSE_x\big(m(x)\big) &= \Big(Bias\big(\hat{m}(x)\big)\Big)^2 + Var\big(\hat{m}(x)\big) \\ &= \Big(E\big[\hat{m}(x)\big] - m(x)\Big)^2 + \Big(E\big[\big(\hat{m}(x)\big)^2\big] - \big(E\big[\hat{m}(x)\big]\big)^2\Big). \end{aligned}$$

By doing so, we can see that by minimising the MSE for a given h, we can achieve a trading-off between the bias and the variance term. In order to monitor *global* accuracy of the estimated $\hat{m}(x)$ over x, we would use *mean integrated square error* defined by

$$MISE\big(\hat{m}(x)\big) = E\int \big(\hat{m}(x) - m(x)\big)^2 dx,$$

which can also be split up into bias and the variance parts (Silverman, 1986, Chapter 3).

Let us assume that the kernel function K is satisfying:

- $\int K(x)dx = 1$,
- $\int xK(x)dx = 0$ and
- $\int x^2K(x)dx = \sigma_K^2 \neq 0$;

that the mean function $m(x)$ has continuous derivatives, $m(x)\colon [0,1] \to \mathbb{R}$, in particular $m^{(p+2)}(x)$, and that the bandwidth h is a non-random sequence of positive numbers that depends on the sample size n such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. By introducing a notation $R_K = \int K(x)^2dx$, for any square integrable function, we can state the standard asymptotic results (see Silverman, 1986, Chapters 3 and Wand and Jones, 1995, Chapter 5):

$$Bias\big(\hat{m}(x)|x_1,\ldots,x_n\big) = \frac{1}{2}h^2\sigma_K^2 m''(x) + o(h^2), \tag{6.3}$$

and

$$Var\big(\hat{m}(x)|x_1,\ldots,x_n\big) = \frac{1}{nh}\sigma_K^2 R_K + o\big((nh)^{-1}\big). \tag{6.4}$$

Equation (6.3) shows that bias depend on x through $m''(x)$, which means that if the function $m(x)$ is close to be linear at x then $m''(x)$ will be very small and close to

zero, while on the other hand if $m(x)$ has high curvature at x then $m''(x)$ would cause local linear fit to produce more biased estimate. From those results we can also see that as the bandwidth becomes larger the bias gets bigger as $O(h^2)$, while the variance decrease as $O\big((nh)^{-1}\big)$.

Wand and Jones (1995, pg. 22) point out that the optimal, *variance-bias trade-off* bandwidth h can be found by differentiating $MISE = Bias^2(\hat{m}(x)) + Var(\hat{m}(x))$ with respect to h and equating the result to 0 giving

$$\hat{h}_{AMISE} = \left(\frac{\sigma_\epsilon^2 R_K}{n\sigma_K^4 R(m'')} \right)^{1/5}, \tag{6.5}$$

where $R(m'') = \int \big(m''(x)\big)^2 dx$, measures the total curvature of the function $m(x)$ and σ_ϵ^2 is a residual variance. Considering this, we can see that for the function $m(x)$ with little curvature we would have a larger optimal bandwidth than for those functions that are more wiggly, with large $R(m'')$ that would cause smaller optimal bandwidth. This seems an easy way of finding the optimal bandwidth, but in practice it is impossible to use considering that we do not know $m(x)$, thus $R(m'')$ is unknown. The only solution to the problem would be to estimate $R(m'')$.

Looking at (6.3), we can also see that the order of the polynomial being locally fitted would determine the order of the bias of $\hat{m}(x)$ for a given bandwidth h. Wand and Jones (1995, Chapter 5) show that for reasonably smooth regression functions, the asymptotic performance of $\hat{m}(x)$ improves for higher orders of the local polynomial fit, although one should be aware that the variance of the estimator, $Var\big(\hat{m}(x)\big)$, would increase for higher value of p. Fitting a polynomial of higher order would possibly reduce the bias, but at the same time, it would increase the variability by introducing more local parameters.

Another question concerns the choice between the even and the odd p order of the local fit. Fan and Gijbels (1996, Capter 3.2) explain the theoretical difference between fitting an odd degree local polynomial and even degree local polynomial. They show that odd order polynomial provides asymptotic bias that has a simpler structure than the even order polynomial. Odd degree polynomial fits have approximate bias depending on x only through $m^{(p+1)}(x)$ and therefore we can observe the error of the

p^{th} degrees polynomial fit by obtaining the $(p+1)^{th}$ derivative of $m(x)$.

Local polynomial smoothing has one very important advantage over other commonly used kernel smoothing estimators, such as Nadaraya-Watson for example and that is the absence of boundary effect. The asymptotic bias at the boundary is of the same order as in the interior without the requirement of using a specific boundary kernel. Because of its appealing bias and boundary properties that would provide, Wand and Jones (1995, pg. 126) suggest odd degree polynomial of either $p = 1$ or $p = 3$ to be used.

6.3 Choosing the Smoothing Parameter

Practical implementation of any kernel smoothing method requires appropriate choice of the smoothing parameter. The choice of the bandwidth would determine the quality of the estimator, and therefore this needs to be done with lots of care.

We know that the optimal bandwidth minimizes MISE and it could be approximated using asymptotic expressions, so that the exact optimal bandwidth is given by (6.5). Unfortunately, we would not be able to use this approximation directly because of the obvious reason we have already pointed out: the expression involves the unknown mean function, $m(x)$, which we are trying to estimate. **Plug-in bandwidth selectors** overcome this problem by "*plugging in*" the estimate of the second derivative of the regression functiom required in the equation (6.5) of the asymptotically optimal bandwidth

$$\hat{h}_{PI} = \left(\frac{\hat{\sigma}_\epsilon^2 R_K}{n \sigma_K^4 \hat{R}(m'')} \right)^{1/5}. \tag{6.6}$$

Ruppert et al. (1995), provide a strategy for plug-n bandwidth selection in which they suggest a 'rule-of-thumb' estimate of the regression function $m(x)$ to be used, based on some parametric fit. Only for the true regression function close to the parametric fit this approach would work reasonably well. In case when the function $m(x)$ is too wiggly, rather than using ordinary least squares to the whole set of data, data can be partitioned and higher order polynomial (suggested $p = 4$) can be used for each subset of data. A draw back to this *blocked quartic fit* approach is that it

requires a choice of the optimal number of blocks N. By using Mallows's C_p (1973) solution to the problem, they find that for relatively smooth regression functions this choice is not so critical and as an optimal number of blocks $\hat{N} = 5$ can be selected, and for the more unstable function higher number of $\hat{N}$ should be used. This method is implemented in `R` as a function `dpill` in the `KernSmooth` package.

Although the combination of block quartic fits and Mallows's C_P provide a relatively easy way of getting the initial estimate of $m(x)$, other possibilities can be considered. Fan and Gijbels (1996, Section 4.2) describe how a similar approach can be used for cases of the mean function with many oscillations for which variable bandwidth would be a better choice

$$\hat{h}(x_i)_{PI} = \left(\frac{\hat{\sigma}_\epsilon^2 R_K}{n f(x) \sigma_K^4 \hat{R}(m'')} \right)^{1/5}. \tag{6.7}$$

assuming that $x_1, \ldots x_n$ are sampled from a $f(x)$ continuous density. They suggest a global fit polynomial of order $p = 3$ to be fitted, that enables the estimation of the second derivative for $m(x)$, which at the same time allows for the certain flexibility in estimating the curvature. The unknown σ_ϵ^2 required in (6.7) can also be easily estimated using this parametric fit. Within the `R` package `locpol`, this approach has been implemented through function `pluginBw`.

Another possible way of selecting the smoothing parameter h is by using **Cross-Validation** approach. In order to define a suitable level of smoothing that would provide a trade-off between the bias and the variance term we use MISE. Cross-validation provides a selection of smoothing parameter by constructing an estimate of MISE and minimising this over h. The main idea behind cross-validation is to predict each response value y_i from the remainder data. Let $\hat{m}_h(x)$ denote the estimate obtained by using smoothing parameter h, of the regression function $m(x)$. For each i, we use data $\{(x_j, y_j), j \neq i\}$ to estimate the *leave-one-out* regression function $\hat{m}_{h,-i}(x)$. The cross-validation function is then defined as

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{m}_{h,-i}(x_i) \right)^2. \tag{6.8}$$

Bowman and Azzalini (1997) show that

$$E[CV(h)] = \frac{1}{n}\sum_{i=1}^{n} E\big[\hat{m}_{h,-i}(x_i) - m(x_i)\big]^2 + \sigma^2, \tag{6.9}$$

so that the cross-validation bandwidth selector is the one that minimises (6.9), hance

$$\hat{h}_{CV} = \underset{h}{\operatorname{argmin}}\, CV(h).$$

`R` has a number of packages that would obtain h_{CV} calculation, such as Bowman's `sm` package with the function `hcv` and package `locpol` with function `regCVBwSelC` .

Figure 6.1 illustrates a cross-validation curve for simulated random sample of bivariate data $\{x_i, y_i\}$, with $x \in [0,1]$ and $x_i = i/n$ for $i = 1, \ldots, n$, where $n = 100$. The y_i's are scalar response variables and the x_i's are predictor variables. The conditional mean function is $m(x) = E(Y|X = x) = \big(0.25 * \exp(2x)\big)^{1/2}$, so that the regression model is given by is $y_i = m(x_i) + e_i$, where $e_i \sim \mathcal{N}(0, 0.04)$. First graph, shows the value of optimal bandwidth which minimises the curve is $\hat{h}_{CV} = 0.1301039$. Using this cross-validation bandwidth $\hat{h}_{CV}$, a local polynomial estimated with $p = 1$ has been constructed and superimposed on the scatter diagram of the simulated data with the true mean function.

Loader (1999) provides a comprehensive study in which he compares different approaches of obtaining the optimal bandwidth. He reports the strongest arguments in favor of plug-in bandwidth selectors have been based on asymptotic studies, in particular the rates of convergence is achieved much faster than other methods. However, the quality of the plug-in selectors relies heavily on the arbitrarily specified pilot bandwidths and do not perform well when this specification is wrong. He concludes that the plug-in approaches are not as good as cross-validation methods, because only for a sufficiently large sample size will the asymptotic eventually take over, and the plug-in selectors be less variable than the one chosen by a cross-validation method.

6.4 Locally Weighted Polynomial Regression

A bivariate smoother is a procedure for obtaining a smooth curve through a scatter diagram. The most commonly used bivariate smoother is the *lowess* curve, which

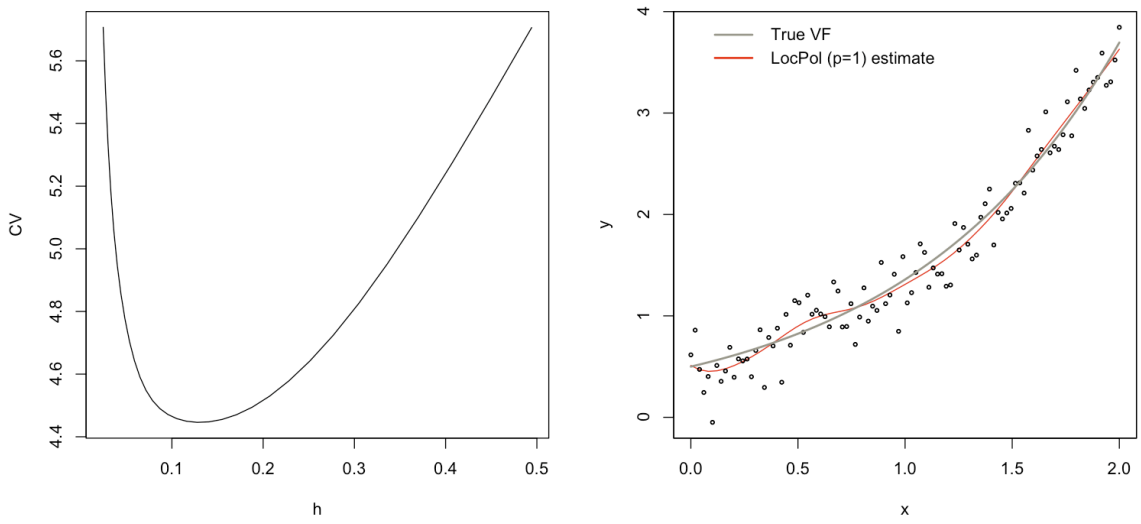


Figure 6.1: *Bandwith Selection Using CV. First plot shows the cross-validation curve for simulated bivariate data of the regression model* $y_i = m(x_i) + e_i$, *with* $x \in [0, 1]$ *and* $x_i = i/n$ *for* $i = 1, \ldots, n$, *where* $n = 100$ *and iid* $e_i \sim \mathcal{N}(0, 0.04)$. *The mean function is given as* $m(x) = \big(0.25 * \exp(2x)\big)^{1/2}$. *The second plot, shows the simulated data with superimposed true mean function* $m(x)$, *showing as the gray line, and a local polynomial of* p *order* 1 *estimate,* $\hat{m}(x)$, *showing as the red line.*

name is an acronym that represents the notion of locally weighted regression function fitting procedure that provides a smooth curve. The algorithm *lowess* for robust locally weighted scatter plot smoothing, was proposed by Cleveland (1979) as an outlier resistant method based on local polynomial fits.

Let us assume that we have bivariate data $\{x_i, y_i\}$, with $x \in [0, 1]$ and $x_i = i/n$ for $i = 1, \ldots, n$. The *lowess* smoothing procedure has been designed to accommodate this data for which

$$y_i = m(x_i) + e_i,$$

where $m(x)$ is a conditional smooth function $m(x) = E(Y|X = x)$, and the iid random error $e_i \sim \mathcal{N}(0, \sigma^2)$. Let us focus on evaluating the regression function at a particular x_0 value, by using a p^{th} order weighted polynomial regression of y on x,

$$y_i = b_0 + b_1(x_i - x_0) + b_2(x_i - x_0)^2 + \cdots + b_p(x_i - x_0)^p + e_i,$$

in which the observations are weighted in relation to their proximity to x_0. Any kernel function can be used (see Table 3.1), but a common weight function used for

lowess is the *tricube*

$$W(x) = \begin{cases} (1-|x|^3)^3 & \text{for} \quad |x| < 1, \\ 0 & \text{for} \quad |x| \geqslant 1. \end{cases} \tag{6.10}$$

The weight, W, is centered at x_0 and it becomes zero at the h^{th} nearest neighborhood point of x_0. By using $W(x)$ as a wight function that decreases for nonnegative x we are able to obtain the wights $w_K(x_0)$ that are decreasing as the distance of x_i from x_0 increase. This is the initial, locally weighted regression estimator that enables the calculation of the residuals $\rho_i = y_i - \hat{y}_i$, which enables a different set of weights to be define. Those *robustness* weights, ρ_i, provide smaller weights to large residuals and larger weights to the smaller residuals. New fitted values are then computed, but with ρ_i that replace $w_K(x_i)$ from the initial model fit. The computation of new weights and new fitted values is now repeated iteratively, several times. The whole process of fitting the initial model and the iterations is referred to as *robust locally weighted regression*, known as *lowess*. Cleveland suggests $p = 1$ order polynomial to be used as it provides a good balance between the easy computation and the flexibility required to reproduce patterns in the data.

6.5 Example 4: Using Non-Parametric Regression Models for Resampling Schemes

To assess how good the regression methods explained in this section would perform in the context of our resampling scheme, we will go back to estimating the 50^{th} ($\tau = 0.5$) quantile function of the homoscedastic, location-shift model (M1) and the heteroscedastic, location-scale shift model (M2):

$$\begin{aligned} \text{Model 1} \quad &: \quad y_i = 2 + 5x_i + e_i \quad \text{and} \\ \text{Model 2} \quad &: \quad y_i = 2 + 5x_i + \sigma(x_i)e_i \end{aligned}$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$. The error, $\{e_i\}$, is iid from tree different distributions: $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$, and the variance is given by $\sigma^2(x_i) = 1 + 4x_i$.

For each of the six simulation studies, we use five hundred realisations ($R = 500$) with a thousand bootstraps ($B = 1,000$). To make an informative comparison, we judge the performance of the used methods by the average lengths (L) and coverage probability (C) of the computed 95% confidence intervals (CI) of the 500 realisations for each of the coefficients in the two given models. We also examine the variability of the model parameters estimates by the *ratio of the determinants* of the estimated covariance matrixes of $\boldsymbol{b}$

$$Covariance\ Ratio = \frac{\det\left[\sigma^2_{boot}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]}{\det\left[\sigma^2_{qr}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]} = \frac{\sigma^2_{boot}}{\sigma^2_{rq}}, \tag{6.11}$$

where σ^2_{boot} is the error variance when using the bootstrap estimatation and σ^2_{qr} is the error variance obtained from the original quantile regression function estimate. A value of 1 would indicate low variability of the estimates obtained by the applied bootstrapping method and indicate that the bootstrapping method is performing relatively well. A ratio value much greater than 1 suggesting bootstrapping method is not performing well and the inferences about the estimated quantile parameters would not be deemed as reliable. In the heteroscedastic case the *ratio of the determinants*

$$Covariance\ Ratio = \frac{\det\left[\sigma^2_{boot}(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}\right]}{\det\left[\sigma^2_{qr}(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}\right]}, \tag{6.12}$$

where

$$\boldsymbol{V} = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_n \end{bmatrix}, \quad \text{with } V_i > 0 \text{ for } i = 1, \ldots, n,$$

is more complex because the *non-identity diagonal matrix*, $\boldsymbol{V}$ need not be the same in both estimates, therefore the determinant of $(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}$ would not cancel top and bottom. However, if the ratio is averaged over a number of sampling runs then if the bootstrapping method is effective we would reasonable expect the ratio to be close to the value of 1 as above. Hence the practical interpretation in the heteroscedastic case would be as above.

We bootstrap using the kernel smoothing bootstrapping adjusted to have the

same mean and the variance as the data from which is constructed (ksm) and the bootstrapping from the corresponding error distributions: $\mathcal{N}(0,1)$, $t(20)$ and $t(10)$.

In the previous simulation studies we have obtained the best overall results when using *Joint Modelling* (see Section 4.6, Appendix B) and when applying *GLM* and *Double GLM* both with *Tweedie* family with *log* link, for estimating the conditional variance function (see Section 5.6, Appendix C). We judge the performance of the newly introduced non-parametric approaches by comparing their results with the results obtained by the selected methods from the previous studies. The full list of the methods used in this simulation study for estimating the conditional variance function is:

- *glm.Tweedie*
- *dglm.Tweedie*
- *Joint Modelling*
- *Local Polynomial with order* $p = 1$
- *Local Polynomial with order* $p = 3$
- *Lowess*.

Prior to deciding on the final list of methods used for estimating the conditional variance function we have examined bias correction of the local polynomial estimate with order $p = 1$.

$$\hat{m}_{bc}(x) = \hat{m}(x) + \frac{1}{2}h^2\sigma_K^2 m''(x) \tag{6.13}$$

The second derivative, $m''(x)$, was obtained from the fitted local polynomial of order $p = 3$. In the models with iid errors, we were getting occasional bias corrected estimates, $\hat{m}_{bc}$, that were very unstable causing the breakdown of the simulation study. This turned out to be even more of a problem when dealing with the heteroscedastic models. The benefit of what we were trying to achieve theoretically does not exist as the underlying conditional variance function is linear, $\sigma^2(x_i) = 1 + 4x_i$, with zero second derivative that yields zero bias. However, the benefits of using bias corrected

local polynomial with $p = 1$ should be examined in case the underlying function has a non-zero second derivative.

We have considered both approaches, the plug-in and the cross-validation, for selecting the bandwidth parameter h for the local polynomial fits. The estimate of the bandwidth parameter by the plug-in approach was surpassingly smaller than the one obtained by the cross-validation, making the estimated curves very wiggly and noisy. The possible reason for this could lie in the fact that we have a linear regression function and the cross-validation is based on the prediction that in this case, is going to be better when using a bigger bandwidth. Consequently, the cross-validation approach was chosen for this study.

From the previous study given in Capter 4, we have realised that the Joint Modelling approach of using GLM proposed by Nelder and Wedderburn (1972) was providing very good results. Unfortunately, we have also discovered the imbedded weakness of this approach that lies in the use of gamma GLM for modeling dispersion parameter, which produces weights of the mean model. The use of the gamma GLM is not robust. We had experienced this in the both, homosecedastic and heteroscedastic cases with the error from $t(10)$ for which the model fitting procedure failed. The previous study, given in Chapter 5, has revealed that we can mimic gamma GLM, by using GLM with Tweedie family with power two, $p = 2$, which has proven to be a more robust alternative. Combining these findings from the previous studies, we were hoping that by modeling the conditional dispersion using Tweedie GLM with $p = 2$, we could improve the robustness of the Joint Modelling approach. The following code shows the adjustment made in the implementation of the Joint Modelling in R:

```
> lmod<-lm(su~1+x)
> for (i in 1:5) {
lhat<-influence(lmod)$hat
d<-((residuals(lmod))^2)/(1-lhat)
# mv<-glm(d~1+x, family=Gamma(link=log), weights=1-lhat) is replaced by:
mv<-glm(d~1+x, family=tweedie(var.power=2, link.power=0), weights=1-lhat)
```

```
w<-1/fitted(mv)
lmod<-lm(su~1+x, weights=w)
}
```

For the model with the error from $t(10)$, for which the squared residuals appear to be very skewed, this approach was still failing for some of the datasets. Figure 6.2 illustrates one of the cases for which Joint Modelling procedure has failed, caused by the extreme observation identified in the scatter plot (observation 13). Removing this extreme observation enables GML with Tweeidie family of $p = 2$ to be fitted to the rest of the data. Unfortunately, we have the same problem highlighted in Chapter 4, disqualifying this approach from the list of the options used for modeling the conditional variance function when dealing with very skewed data, which is the model with error from $t(10)$.

Figure 6.3 illustrates fitted conditional variance functions for the six models in the Example 4, adopting the approaches as indicated in list above. The full list of the results of the simulation study is given in Appendix D.

6.5.1 Model 1: Homoscedastic, Location-Shift Model

The methods used provide very close parameter estimates to their true values for all three models. The intercept parameter estimate (b_0) is slightly under-estimated and the slope parameter estimate (b_1) slightly over-estimated in all three models by all of the methods used, except for the model with error $\{e_i\}$ from $t(20)$ where the competing bootstrapping methods (*xy*, *pwy*, *mcmb* and *wxy*) slightly under-estimate the slope parameter.

The coverage probabilities of the coefficients in the three models are very competitive between the methods used in the study. The rank-score method produces the shortest lengths of the parameters' confidence intervals, but the coverage probabilities are the smallest making this approach far less attractive. Comparing the results of the remaining methods, for the models with error $\{e_i\}$ for $t(20)$ and $t(10)$ the method that provides shortest average length of the parameters' confidence intervals is Wald

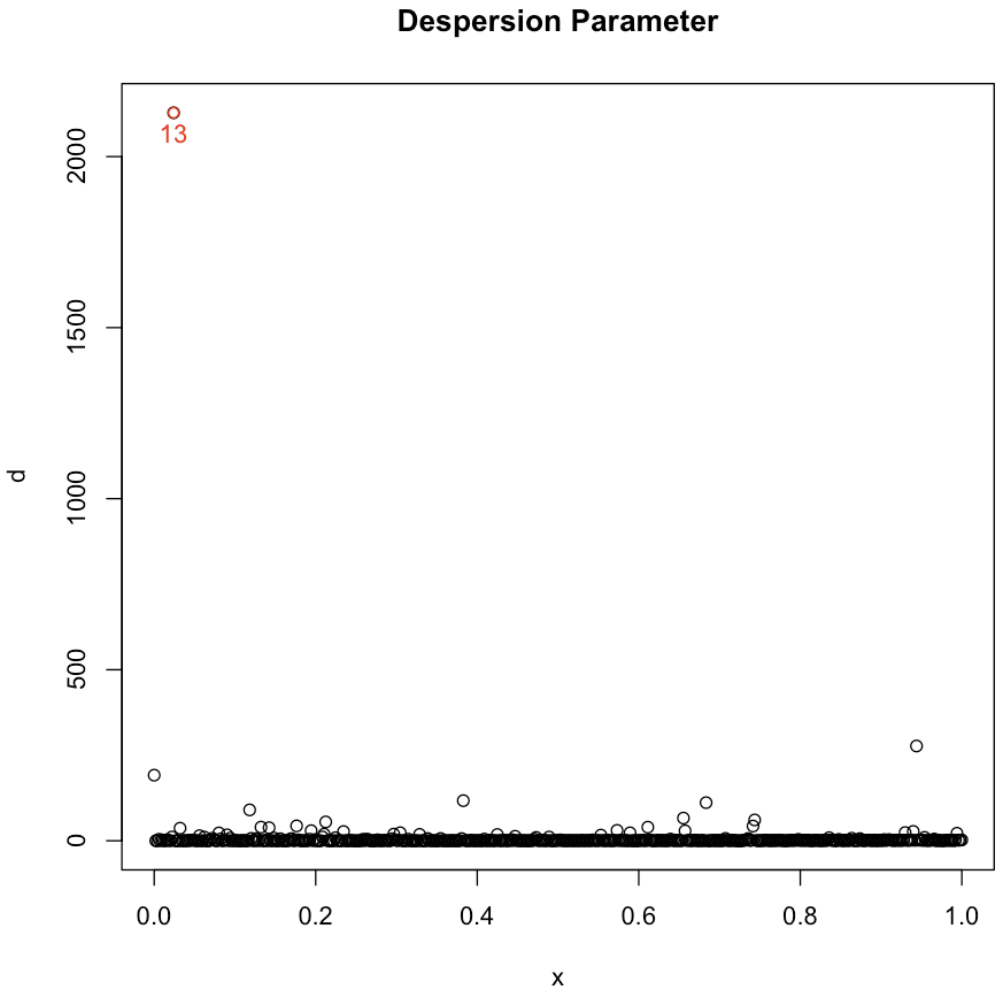


Figure 6.2: *Dispersion Parameter. Figure shows a scatter plot of the dispersion parameter used in the Joint Modelling approach for the homoscedastic model with the error from* $t(10)$.

method, when iid error is assumed (*wiid*). However, the standard error of the mean length confidence interval for the *wiid* is twice of the standard error obtained by the kernel smoothing bootstrapping methods.

When comparing between the two resampling approaches used, only for the model with the error from normal distribution the bootstrapping from the same standardised distribution centered on the estimated quantile, provides shorter parameter confidence intervals. For the other two models the kernel smoothing bootstrapping adjusted to have the same mean and the variance as the data from which is constructed, *ksm* is the preferred one.

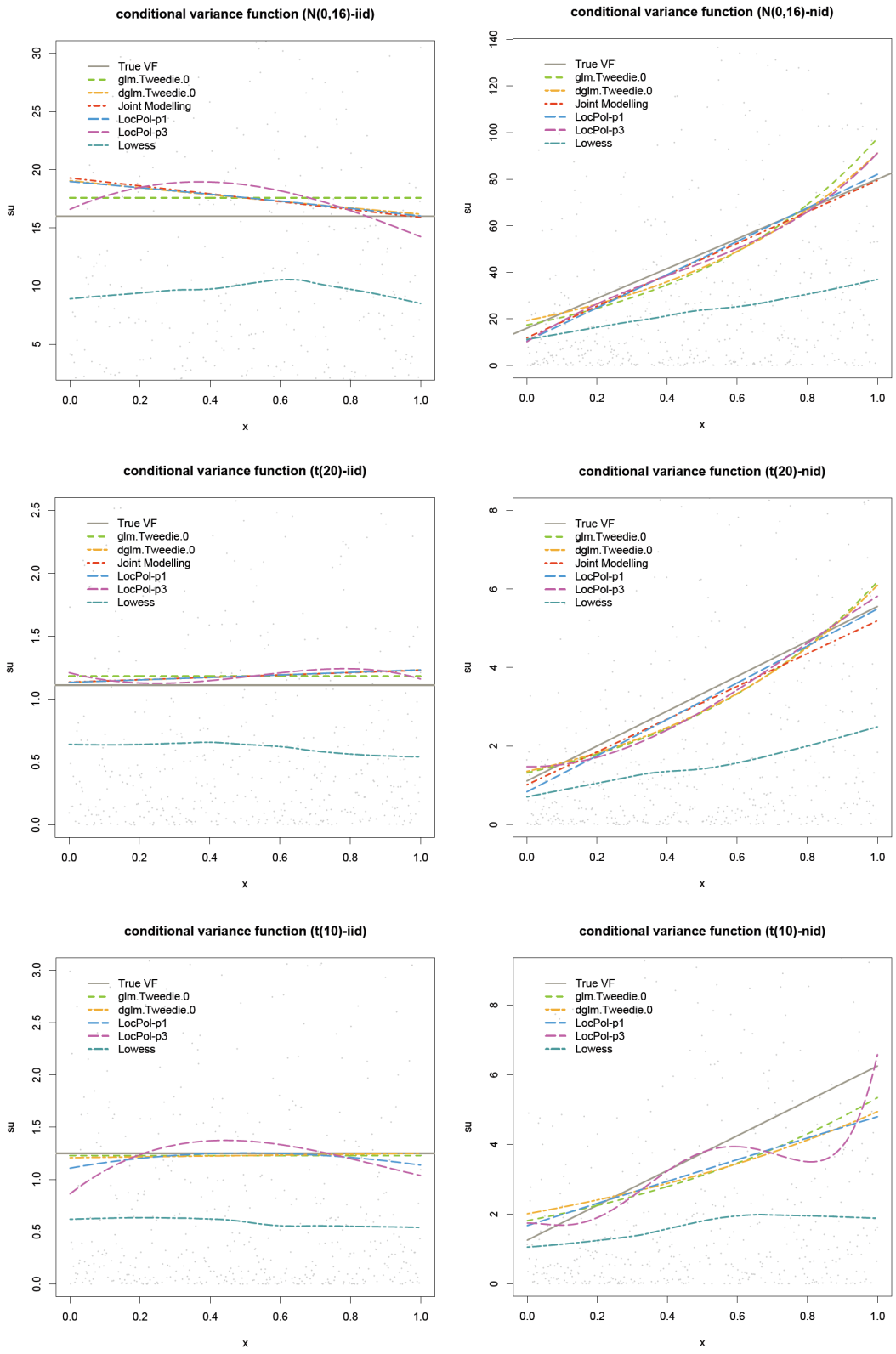


Figure 6.3: *Conditional Variance Function Estimations for Example 4 models. The figures plot the squared centered residuals of the simulated data for the two models given in the Example (M1 and M2) with the errors from the three different distributions:* $\mathcal{N}(0, 16)$, $t(20)$ *and* $t(10)$. *Solid gray line on the plots represents the true variance function. The estimates of the conditional variance functions are given in different colored dashed lines: green, orange, red, turquoise, magenta and royalblue when using the following methods: glm.Tweedie.log, dglm.Tweedie.log, JointModeling, LocPol-p1-res, LocPol-p3-res, Lowess, respectively.*

For the models with error from $\mathcal{N}(0, 16)$ and $t(20)$ modeling the conditional variance function by the Joint Modelling technique and the Tweedie.DGLM provide better results than the non-parametric approaches. Only in the model with the error from $t(10)$ parametric methods used are outperformed by the non-parametric local polynomial model fitting procedures. The combination of applying *lowess* for modeling the conditional variance and resampling from the error distribution, provides the shortest average lengths of the parameters confidence intervals with very small standard error. However, for this combination, the mean coverage probability is between 5% to 10% below the others making it a less attractive combination choice.

6.5.2 Model 2: Heteroscedastic, Location-Scale Model

The parameter estimates provided by all of the used approaches are very close to their true values in all three models. The intercept parameter estimate (b_0) is consistently slightly under-estimated, whilst the slope parameter estimate (b_1) slightly over-estimated in all three models by all methods used, except for the model with error $\{e_i\}$ form $t(20)$ where all of the competing bootstrapping methods (xy, pwy, mcmb and wxy) slightly under-estimate the slope parameter.

The coverage probabilities of the models' coefficients are very high and similar for all of the approaches with the exception to the rank-score, which provides the narrowest lengths of the parameter's confidence intervals, but low coverage probability in comparison to the other methods. When using kernel smoothing bootstrap approaches, the average length of the confidence interval of the intercept parameter is still wider than those obtained by the other competing bootstrapping methods. Looking at the results when using our resampling approaches, the shortest average lengths of the confidence intervals for the estimated intercept parameter b_0 are obtained when resampling from the same error distribution. In terms of the estimated slope parameter, b_1, kernel smoothing bootstrapping approaches perform better than the other bootstrapping methods.

Comparing the results obtained by the kernel smoothing bootstrapping approaches

it appears that if the heteroscedastic model has the error from the normal distribution parametric method for estimating the conditional variance function should be used. In the models with error form $\mathcal{N}(0, 16)$ modeling the conditional variance function by the Joint Modelling technique, the Tweedie.DGLM and the Tweedie.GLM provide better results than the non-parametric approaches. For the models with non-normal error distribution, the local polynomial of order one ($p = 1$) provides the best results, although they are closely followed by the parametric methods used. Like in the case of homoscedastic models, here applying *lowess* for estimating the conditional variace in the combination of resampling from the error distribution provides the shortest average confidence interval lengths, with very small standard error. However, in terms of the coverage probability the percentage is up to 10% lower than those of the other approaches used with the exception to the rank-score method that provides similar results.

Applying six different modeling approaches to the conditional variance the average value of the variance of the standardised residuals when using local polynomial of $p = 3$ and *lowess* is over 1. The standard errors of those averages are very high in comparisons to the others, especially in the case of the local polynomial with $p = 3$, suggesting high variability. Looking at the matrix norms in terms of the ratio of the determinants, we notice higher variability when using those two modeling approaches to estimating the conditional variance. For the model with the normal error from $\mathcal{N}(0, 16)$, when using local polynomial of $p = 3$ the mean ratio of the determinants is extremely high. The probable cause of such a high variability in the model is the high variance of the error term.

6.6 Conclusion

In this simulation study we have modeled the conditional variance function using non-parametric approach of fitting local polynomial models to the data. The advantage offered by the non-parametric methods is that the functional form does not have to be specified, giving more flexibility in modeling the data. Another benefit of using

a non-parametric regression technique is that no assumption has to be made on the distribution of the response, which when dealing with our resampling method could be of a particular interest.

In the study given in Chapter 4, we have recognised the potential of using the Joint Modelling parametric technique as it was providing very good results. By applying the gamma.GLM modeling of the dispersion element, this approach appears not to be robust. To overcome this problem in making the Joint Modelling approach robust, when modeling the dispersion element we have proposed the use of the Tweedie.GLM with power $p = 2$, that acts as a pseudo Gamma.GLM. Unfortunately, as this study has shown, even Tweedie.GLM that has proven to be robust for the models with the error from $t(10)$ fails in the presence of the extremely big outliers (see Figure 6.2), not enabeling us to use the Joint Modelling approach for the models with the error from $t(10)$.

When using non-parametric methods for modeling the conditional variance function of the given quantile regression models (M1 and M2) the local polynomial of order 1 proved to be a better fit than the local polynomial of order 3 or *lowess*. The exception was the homoscedastic model with the error from $t(10)$ for which the narrowest confidence intervals of the coefficients are obtained when modeling the conditional variance function by the local polynomial of order three, $p = 3$. The local polynomial of order 3 has more desirable theoretical properties than the local polynomial of order 1, however, as this study has shown, it provides more flexibility in the model estimate than is needed. This appears to be the case as in the homoscedastic models the ratio of the determinants of the estimated covariance matrixes of b is far higher that the value of 1, indicating high variability in the model estimate when using this method. The robust, non-parametric *lowess* method offers a way around dealing with the extreme observations resulting in the estimated conditional variance function to be perhaps pulled too far down. The mean variance of the standardised residuals is above the value of 1 for both homosecedastic and heteroscedastic models. When using this method, we can achieve relatively good results only if bootstrapping from the error distribution, which in practice is most likely to be unknown.

Slightly over-inflated intercept parameter remains to be a problem that the applied non-parametric methods for estimating the conditional variance function are still not able to address. In the following study we are going to examine the possibilities offered by the second class of variance estimators known as *difference-based estimators* in an attempt that this approach would help to remedy the problem. As a result of this simulation study, we can conclude that if we want to allow for some flexibility in not specifying a functional form of the conditional variance function, then local linear model of order 1 provides a good competitor to the selected parametric method used in this study.

Chapter 7

Difference-Based Variance Function Estimation

7.1 Introduction

Difference-based kernel estimators for the error variance are popular since they do not require the estimation of the mean function. Not like the residual based approach, which smooths squared residuals from the fitted quantile regression model, the difference-based method for estimating the variance function enables the estimate of the variance which is independent of the mean function. This class of variance estimators uses the differences of the combined observations as the response variable with the aim to remove the trend in the mean function, an idea originating in time series analysis (Tong and Wang, 2005).

The previous study in Chapter 6 has revel that when dealing with the homoscedastic models with non-normal errors the use of the local polynomial estimation of the conditional variance function improves the results of the kernel smoothing bootstrapping method for the quantile regression models. In this section we will apply a different non-parametric, difference-based approach for estimating the conditional variance function that would be used in the context of our kernel smooth bootstrapping method. The residuals obtained from the estimated quantile regression function will be used to calculate the differences that would act as the response variable in

the conditional variance model estimation.

7.2 Difference-Based Variance Estimators

Let us consider the following regression model

$$y_i = m(x_i) + \sqrt{V(x_i)}e_i, \quad i = 1, \ldots, n,$$

where $m(x_i)$ and $V(x_i)$ are unknown mean and variance functions respectively. The errors e_i are independent standard normal variables $\mathcal{N}(0, 1)$. Let us also assume that $x \in [0, 1]$ and that $x_i = i/n$. The problem of interest is to estimate the variance $V(x_i)$ in the presence of the unknown mean $m(x_i)$, which plays the role of the nuisance parameter.

The order of the difference-based estimator is defined as the number of related observations involved in calculating a local residual. Rice (1984) proposed the normalised sum of squares first-order differences

$$\hat{V}(x)_R = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \tag{7.1}$$

as an estimator when $V(x) \equiv \sigma^2$ is assumed constant. Gasser (1986) proposed the second-order difference-based estimator

$$\hat{V}(x)_{GSJ} = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \left(\frac{1}{2} y_{i-1} - y_i + \frac{1}{2} y_{i+1} \right)^2. \tag{7.2}$$

General difference-based estimators of an arbitrary order k was introduced by Hall and Marron (1990)

$$\hat{V}(x)_{HKT}(k) = \frac{1}{n-k} \sum_{i=1}^{n-k} \left(\sum_{j=0}^{k} d_j y_{j+1} \right)^2, \tag{7.3}$$

where k a nonnegative integer. The difference sequence d_i, $i = 0, \ldots, k$ is defined such that its elements sum up to zero:

$$\sum_{i=0}^{k} d_i = 0, \tag{7.4}$$

while the sum of squares is 1:

$$\sum_{i=0}^{k} d_i^2 = 1. \tag{7.5}$$

In the previous Chapter 6, we have seen the application of the standard idea, which uses a normalised sum of squared residuals as an estimator of the variance (for example see Hall and Carroll, 1989). For a linear smoothing function the fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{H}_x\boldsymbol{y}$, where $\boldsymbol{H}_x = \big(\boldsymbol{X}^\top\boldsymbol{W}_x\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{W}_x$ and $\boldsymbol{W} = diag\{w_i(x)\}$ is a diagonal kernel weight matrix with diagonal elements $w_i(x) = K\big((x - x_i)/h\big)$. The vector of the residuals can be expressed as $\boldsymbol{U} = (\boldsymbol{I} - \boldsymbol{H}_x)\boldsymbol{y}$ which gives the sum of the squared residuals as $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{y}^\top\boldsymbol{D}_x\boldsymbol{y}$, where $\boldsymbol{D}_x = (\boldsymbol{I} - \boldsymbol{H}_x)^\top(\boldsymbol{I} - \boldsymbol{H}_x)$. This residual-based estimator can be normalised by dividing it with $tr\boldsymbol{D}_x$

$$\hat{\sigma}_D^2 = \frac{\boldsymbol{y}^\top\boldsymbol{D}\boldsymbol{y}}{tr\boldsymbol{D}}. \tag{7.6}$$

Levins makes the remark (2003, Remark 3.5) in which he points out the importance of (7.6) as (7.2) represents a normalised quadratic form

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^\top\boldsymbol{D}\boldsymbol{y}}{n - k}, \tag{7.7}$$

with the matrix $\boldsymbol{D} = \tilde{\boldsymbol{D}}^\top\tilde{\boldsymbol{D}}$ where

$$\tilde{\boldsymbol{D}} = \begin{bmatrix} d_0 & \dots & d_k & 0 & \dots & 0 \\ \vdots & \ddots & & \ddots & & \vdots \\ \vdots & & \ddots & & \ddots & \vdots \\ 0 & \dots & 0 & d_0 & \dots & d_k \end{bmatrix} \tag{7.8}$$

is a banded matrix with the $k + 1$ bands.

The k^{th}-order variance estimators (7.1), (7.2) and (7.3) represent the scaled global averages of the squared differences

$$\Delta_{k,i}^2 = \sum_{j=0}^{k} (d_j y_{j+i-\lfloor\frac{k}{2}\rfloor})^2,$$

where $\{d_j\}$ is a difference sequence as defined by (7.4) and (7.5) and $i = \lfloor\frac{k}{2}\rfloor + 1, \dots, n + \lfloor\frac{k}{2}\rfloor - k$. The differences (7.1) represent 1-dependent sequence, by using

first-order difference, Δ_1, we can obtain pseudo residuals $r_i = \frac{1}{2}(y_i - y_{i-1})^2$, which we can also write as

$$r_i = \Delta_{1,i}^2 = \sum_{j=0}^{1}(d_j y_{j+1})^2, \quad \text{where } d_0 = \frac{1}{\sqrt{2}}, \text{ and } d_1 = -\frac{1}{\sqrt{2}}.$$

In the case of heteroscedasticity where the variance is not constant, but a function of the covariate x, we should consider local estimate of $V(x)$ in the neighborhood of x. By applying local polynomial regression (see Section 6.2) to the k^{th}-order calculated differences $\Delta_{k,i}$ for all $i \in \{1, \ldots, n-k\}$ the local average of those differences can be obtained. This would define the variance estimator as the weighted local average of the squared pseudoresiduals of order k that would allow for non-constant variance

$$\hat{V}(x) = \frac{\boldsymbol{y}^\top \boldsymbol{D}(x)\boldsymbol{y}}{tr(\boldsymbol{D}(x))}. \tag{7.9}$$

Matrix $\boldsymbol{D}(x)$ is a $n \times n$ positive, semi-definite symmetric matrix, $\boldsymbol{D}(x) = \tilde{\boldsymbol{D}}^\top(x)\tilde{\boldsymbol{D}}(x)$, where

$$\tilde{\boldsymbol{D}}(x) = \begin{bmatrix} d_0\sqrt{K(\frac{x-x_i}{h})} & \ldots & d_k\sqrt{K(\frac{x-x_i}{h})} & 0 & \ldots & 0 \\ \vdots & \ddots & & \ddots & & \vdots \\ \vdots & & \ddots & & \ddots & \vdots \\ 0 & \ldots & 0 & d_0\sqrt{K(\frac{x-x_{n-k}}{h})} & \ldots & d_k\sqrt{K(\frac{x-x_{n-k}}{h})} \end{bmatrix} \tag{7.10}$$

is a $(n-k) \times n$, banded matrix with $2k+1$ bands for an estimator of order k.

A variance estimator $\hat{V}(x)$ of order k can be obtained by applying local polynomial regression estimation using $\Delta_{k,i}^2$. In the case of fitting the local polynomial of $p = 0$ we would have the Nadaraya-Watson estimator

$$\hat{V}(x_i) = \frac{\sum_{i=1}^{n-k} \Delta_{k,i}^2 K(\frac{x-x_i}{h})}{\sum_{i=1}^{n-k} K(\frac{x-x_i}{h})}. \tag{7.11}$$

and we have

$$\hat{V}(x_i) = \hat{a}_0, \tag{7.12}$$

where

$$(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p) = \underset{\{a_0,a_1,\ldots,a_p\}}{\operatorname{argmin}} \sum_{i=\lfloor \frac{k}{2} \rfloor + 1}^{n+\lfloor \frac{k}{2} \rfloor - k} [\Delta_{k,i}^2 - a_0 - a_1(x - x_i) - \cdots - a_p(x - x_i)^p]^2 \times K\left(\frac{x - x_i}{h}\right). \tag{7.13}$$

Brown and Levine (2007) in their study on optimal convergance of the difference based method for variance estimation also establish its basic asymptotic properties.

When applying local polynomial regression we assume that the data is independent. By using the squared differences as the response variable in the local polynomial regression modelling the assumption of independent data is compromised as the squared pseudoresiduals, $\Delta_{k,i}^2$ construct a k-dependent sequence. As a result the standard asymptotic theory of obtaining bias and variance can not be applied, which poses a difficulty when selecting a bandwidth parameter for the local polynomial fit.

7.3 Bandwidth Selection

Despite its popularity and the development in terms of its application (for some recent ones see Müller et al. 2003; Tong and Wang 2005) the issue of the bandwidth selection has not been adequately addressed for the difference-based variance estimation method. In a recent paper Levine (2006) advocates the use of the cross-validation approach for bandwith selection.

Let us assume that $m(x) \in C^1[0,1]$ and $V(x) \in C^2[0,1]$. The kernel function is a second-order function $K(x) : [0,1] \to \mathbb{R}$ satisfying $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx = \sigma_K^2 \neq 0$. As in the previous chapter, we will adopt standard notation $R_K = \int K^2(x)dx$. For the estimator (7.11) Levine (2003) provide the exact optimal bandwidth

$$h_o = \left[\frac{CR_K \int V^2(x)dx}{2n\sigma_K^2 \int \left(V''(x)\right)^2 dx} \right]^{1/5}, \tag{7.14}$$

where

$$C = 2\left(1 + 2\sum_{l=1}^{k} \left(\sum_{j=0}^{k-1-l} d_j d_{j+l} \right)^2 \right) \tag{7.15}$$

is a constant that depends on the chosen sequence $\{d_i\}$, as for any given value of k there is a difference sequence that minimisis this constant. Hall et al. (1990) show that asymptotic variance, but not the bias of the estimator can be affected by the choice of the difference sequence $\{d_i\}$.

Using a **plug-in** approach for estimating the unknown bandwidth h_o requires

a difficult estimation of another two quadratic functions in (7.14): $\int V^2(x)dx$ and $\int \left(V''(x)\right)^2 dx$. This difficulty makes the estimation of the bandwidth parameter a more complex problem than the estimation of the variance function itself. In terms of its simple application, the **cross-validation** (CV) approach for estimating the bandwidth parameter seems a more appealing alternative. However, as already pointed out, in the case of the difference based variance estimation the squared pseudoresiduals, $\Delta_{k,i}^2$ construct a k-dependent sequence, which creates an obstacle for application of the cross-validation approach which assumes uncorrelated data. Chu and Marron (1991) point out that in the case of *leave-one-out* CV when the data is positively correlated, cross-validation will tend to produce small bandwidths which will result in under-smoothed estimates of the regression function. On the other hand, if the observations are negatively correlated, then cross-validation will tent to produce large bandwidths which will result in over-smooth estimates of the regression function.

7.3.1 Cross-Validation for Correlated Data

Levine (2006) proposes a **K-fold cross-validation for correlated data** that can be adopted for the implementation of the bandwidth selection for the difference-based variance estimator. The standard K-fold CV in which $1 < K < n$, is an alternative to the *leave-one-out* CV ($K = n$) that reduces the computational time without diminishing the characteristics of the original data. Anthony and Holden (1998) indicate that leave-one-out CV potentially produces over-fitting which results in poor generalisation error, which is the reason why they advocate the use of the K-fold CV. Considering this and the fact that the estimator (7.11) has the variance that is decreasing as k rises Levine proposes a version of K-fold cross-validation similar to that for independent data.

Cross-validation loss function approximates integrated squared error (ISE)

$$ISE = \int \left(\hat{f}(x) - f(x)\right)^2 dx$$

and the expectation of cross-validation loss function approximates the mean integrated squared error (MISE), that is a basis of any cross-validation type method

$$MISE = E \int \left(\hat{V}(x) - f(x)\right)^2 dx, \tag{7.16}$$

which depends on the unknown function $V(x)$ by some data-depenent quantity. As a way for estimating (7.16) Levine (2006) suggests the data $\{x_i, y_i\}$ to be partitioned into K approximately equal and disjoint subsets, where each of the created subsets would have k_j pairs, such that $\sum k_j = n$. Let us denote *the validation data* pairs in the j^{th} subset as $\{\breve{x}_{(j,i)}, \breve{y}_{(j,i)}\},\ i = 1, \ldots, n_{k_j}$, where $\breve{x}_{(j,i)}$ are arranged in ascending order, $\breve{x}_{(j,i)} \leqslant \breve{x}_{j+1,i},\ i = 1, \ldots, k_j - 1$. Also, let us denote *the training data* pairs in the j subset as $\{x_{(j,i)}, y_{(j,i)}\},\ i = 1, \ldots, n_{k_j}$, where $x_{(j,i)}$ with the $x_{(j,i)}$ also arranged in ascending order. According to this notation, we can denote the pseudo residuals from $\{x_{(j,i)}, y_{(j,i)}\}$ and $\{\breve{x}_{(j,i)}, \breve{y}_{(j,i)}\}$ as $\Delta_{(j,i)}$ and $\breve{\Delta}_{(j,i)}$, respectively. Adopting this notation, the estimator derived from the j^{th} subset $\{x_{(j,i)}, y_{(j,i)}\}$ when using bandwidth h and squared pseudoresidual $\left(\Delta_{(j,i)}\right)^2$ can be denoted as $\hat{V}_{h,j}$. By defining

$$CV_j(h) = \sum_{l=1}^{k_j} \left(\left(\breve{\Delta}_{(j,l)}\right)^2 - \hat{V}_{h,j}(\breve{x}_l) \right)^2$$

the cross-validation criterion is

$$CV(h) = \frac{1}{n} \sum_{j=1}^{K} CV_j(h), \tag{7.17}$$

which allows the choice of the optimal bandwidth

$$h_{CV} = \underset{h \in H}{\operatorname{argmin}}\, CV(h), \tag{7.18}$$

where H is the finite grid. Minimising this $CV(h)$ criterion (7.17) is equivalent to minimising the MISE of the variance estimator $\hat{V}(x)$ in (7.16). Levine (2006) also conduct a study of the performance of the proposed method depending on the choice of K. He suggests that unless the sample size is quite small $K = 10$ is a good choice.

In general, the CV method assumes independent data, thus the validation set and the training set need to be independent from each other. Adopting the CV method to the k order differences, which presents a k-dependent sequence, means that a

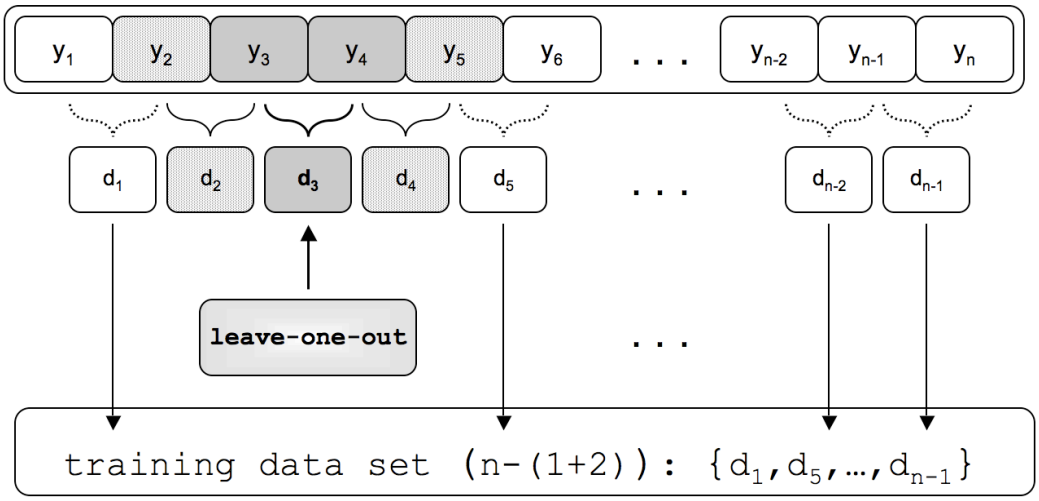


Figure 7.1: *Leave-One-Out+2k Cross Validation. Figure illustrates the calculation of the* 1^{st} *order* ($k = 1$) *differences and the selection of the training data used used in the leave-one-out+2k cross-validation method applied to the* 3^{rd} *difference. The prediction is calculated for* d_3 *using training data set which includes* $\{d_1, d_5, \ldots, d_{n-1}\}$.

way of diminishing the existing correlation needs to be found. For equally spaced data and small value of k, such as $k = 1, 2$ or 3, that implies a relatively low-depth correlation between the pseudoresiduals Δ_j^2, we propose the use of the **leave-one-out+2*k* cross-validation** method. When applying the *leave-one-out*+2k CV from the training data we pull out the adjacent neighboring differences that are correlated to the difference being predicted. Figure 7.1 illustrates the application of the $(1+2k)$ cross-validation method applied to the 1^{st} order ($k = 1$) difference sequence. The prediction is calculated for difference d_3 using the training data set $\{d_1, d_5, \ldots, d_{n-1}\}$, which does not include the two neighboring differences d_2 and d_4 to which d_3 is correlated. Removing the correlated differences from the training data sets enables the independence between the data used in the CV procedure.

The leave-one-out+2k CV offers an easy implementaion of the CV when dealing with the k-dependent sequence data. The drawbacks are that it is computationally time consuming and that for increased k it becomes wasteful.

For the classes of estimators defined by (7.11) by increasing k from 1 to ∞, Levine (2003) establishes that it can achieve at most around 33% reduction in variance. Considering this there is a serious doubt on whether higher-order estimators will

yield further significant improvement.

7.4 Example 5: Using Difference Based Variance Function Estimation for Resampling Schemes

Let us go back to the problem of estimating the 50^{th} ($\tau = 0.5$) quantile function of the homoscedastic, location-shift model (M1) and the heteroscedastic, location-scale shift model (M2):

$$\begin{aligned} \text{Model 1} \quad &: \quad y_i = 2 + 5x_i + e_i \quad \text{and} \\ \text{Model 2} \quad &: \quad y_i = 2 + 5x_i + \sigma(x_i)e_i \end{aligned}$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$. The error, $\{e_i\}$, is iid from four different distributions:

i) $\mathcal{N}(0, 0.04)$,

ii) $\mathcal{N}(0, 16)$,

iii) $t(20)$ and

iv) $t(10)$.

For the purpose of the comparison of the results for the models with the error from the normal distribution with the high variability ($\sigma = 4$) in this study we also consider the error from the normal distribution with relatively small variability ($\sigma = 0.2$). For model two (M2), this time we use two different variance functions:

a) $\sigma^2(x_i) = 1 + 4x_i$ and

b) $\sigma^2(x_i) = \frac{1}{4}\exp(2x_i)$.

For each of simulation studies, we use five hundred realisations ($R = 500$) with a thousand bootstraps ($B = 1,000$). As in the previous studies, to make an informative comparison, we judge the performance of the used methods by the average lengths

(L) and coverage probability (C) of the computed 95% confidence intervals (CI) of the 500 realisations for each of the coefficients in the two given models. The variability of the model parameters estimates is examined by the *ratio of the determinants* of the estimated covariance matrixes of $\boldsymbol{b}$ given by (6.11) for homoscedstic Model 1 and by (6.11) for heteroscedastic Model 2. In practice, the error distribution is most-likely to be unknown, in this study we will not bootstrap from the corresponding error distributions, but rather focus only on the kernel smoothing bootstrapping adjusted to have the same mean and the variance as the data from which it is constructed (*ksm*).

In the previous studies, methods used for estimation of the conditional variance function that have provided best results when used in the context of the kernel smoothed bootstrapping are double generalised linear model from Tweedie family with the log link and the residual-based approach using a *local polynomial of order* $p = 1$. The results obtained using those two methods for estimating the conditional variance function are going to be compared with the results provided when using a difference-based approach.

This study includes a nonlinear variance function $\sigma^2(x_i) = 0.25\exp(2x)$ with a continuous second derivative. When using the local polynomial fits of order one, we have considered bias correction (6.13). Unfortunately, the estimates of the second derivatives proved to be unstable, making bias correction etimates of the fitted local polynomials of order one inaccurate.

To illustrate this, we run a separate study for which we define a *loss function* that characterises the deviation of the second derivative estimate function from the true second derivative function at a point. We define the squared error loss function $L(x_i) = \left(\hat{f}''(x_i) - f''(x_i)\right)^2$, where $\hat{f}''(x_i)$ is estimated second derivative of the function $f(x)$ at a given x_i and $f''(x_i)$ is the true value of the second derivative of the function $f(x)$ at a given point. This allows the definition of the global measure $L(\boldsymbol{x}) = (1/n)\sum_{i=1}^{n} L(x_i)$. The following choices of the functions are considered:

i) $f_1(x) = \sin^3(2\pi x^3)$ and

ii) $f_2(x) = (\frac{1}{4}\exp(2x))$.

The first is a wiggly sinusoid function with high oscillation, the second being a smooth exponential function. Figure 7.2 illustrates the second derivatives of the two functions.

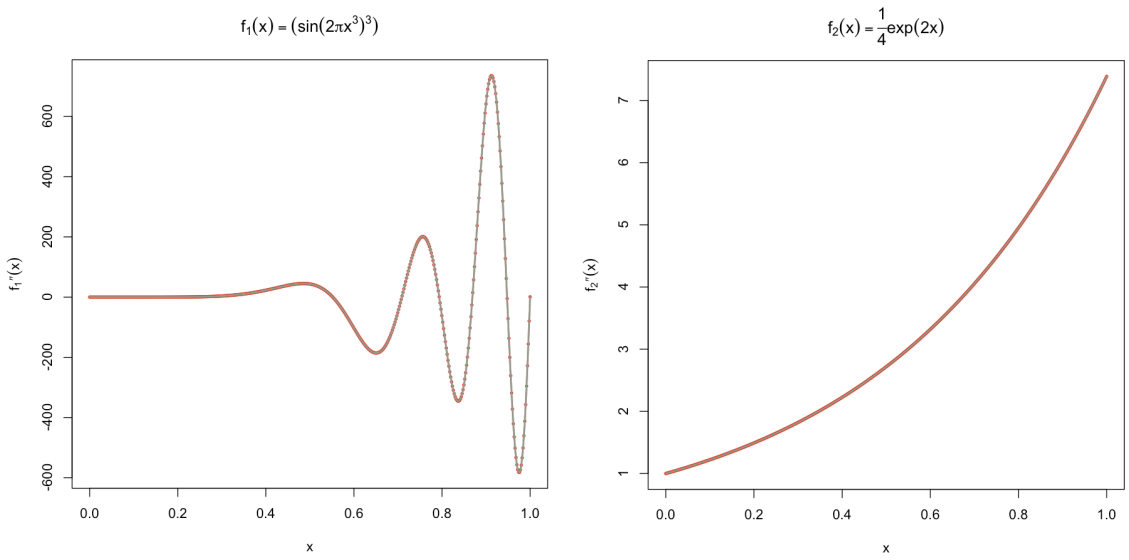


Figure 7.2: *Second derivatives. Figures illustrate the second derivatives of the two functions* $f_1(x) = \sin^3(2\pi x^3)$ *on the left plot, and* $f_2(x) = (\frac{1}{4}\exp(2x))$ *on the right plot.*

We consider a homoscedastic and a heteroscedastic problem for each of the functions with the error from the normal distribution, $e_i \sim \mathcal{N}(0, \sigma^2)$, with sigma increasing sequentially by 0.25, starting with 0.25 and going up to 4. We also look at an equally spaced design where $x_i = 1/n$, $i = 1, \ldots, n$, for $n = 500$ and perform 100 simulation for each function and choice of σ^2. The derivatives are estimated using a local polynomial fit for which a *leave-one-out* cross-validation bandwidth has been estimated and used. Table 7.1 shows the results of the study. It is interesting to note that the estimate of the second derivative of the sinusoid function $(f_1(x))$ becomes more stable in the heteroscedastic setting than it is in homoscedastic, but there is a high variation in the estimate irrespective of the size of the variance. Regarding the exponential function $(f_2(x))$ the instability of the second derivative estimate becomes more obvious as the variance σ^2 increases. We were concerned about the appearance of such randomly varying $L(\boldsymbol{x})$ with the change in σ^2 and with the closer examination

of the $L(x_i)$ values we have noticed that for the certain data sets $L(\boldsymbol{x})$ result in a very large values. What becomes evident is that with the increasing value of σ^2 the estimation of the second derivative becomes unstable for both the homoscedastic and the heteroscedastic cases.

	Homoscedastic		Heteroscedastic	
σ	$\sin^3(2\pi x^3)$	$\frac{1}{4}\exp(2x)$	$\sin^3(2\pi x^3)$	$\frac{1}{4}\exp(2x)$
0.25	66,725.97	34.45	38,199.85	23.29
0.50	137,701.13	28.45	38,274.43	270.57
0.75	153,518.45	29.81	38,288.40	275.10
1.00	142,006.16	114.52	38,701.15	559.09
1.25	95,802.25	390.48	38,915.68	125.07
1.50	98,682.76	341.43	38,525.57	2,368.00
1.75	136,753.56	1,776.63	38,434.36	3,925.72
2.00	220,359.47	681.60	39,660.03	27,629.29
2.25	116,521.52	1,847.31	38,562.13	4,301.84
2.50	82,378.05	642.70	39,114.68	5,136.12
2.75	576,229.26	623.71	39,621.58	19,972.91
3.00	68,394.67	37,228.52	39,407.02	11,676.57
3.25	62,071.40	844.69	46,433.81	29,408.32
3.50	97,655.68	3,898.59	51,892.76	18,164.71
3.75	68,715.03	1,243.71	39,032.34	21,815.66
4.00	86,176.05	3,285.09	38,808.41	368,880.73

Table 7.1: Loss Function. Table shows average values of $L(x_i) = \left(\hat{f}''(x_i) - f''(x_i)\right)^2$ for the two functions: $f_1(x) = \sin^3(2\pi x^3)$ and $f_2(x) = (\frac{1}{4}\exp(2x))$. Two different models are considered: homoscedastic and heteroscedastic with the error from $\mathcal{N}(0, \sigma^2)$.

Considering that the estimate depends upon the bandwidth parameter h used for the local polynomial fit, we additionally examine the change in the loss function $L(\boldsymbol{x})$ when bandwidth h changes sequentially starting from 0.25 and going up to 2.00 in the steps of 0.25. Again we perform 100 simulations for each combination of h and σ^2 for the two functions for the heteroscedastic setting.

The average value of the loss function changes very little in the case of the sinusoid function f_1 (see Table 7.2). If we observe the corresponding standard errors we notice that they are increasing with increasing σ^2. However, the standard error does not appear to be affected very much by the changes in the bandwidth parameter h. For a given value of σ, the standard error of $L(\boldsymbol{x})$ is the highest for the smaller values of h,

	$\boldsymbol{\sigma}$							
$\boldsymbol{h}$	0.5	1	1.5	2	2.5	3	3.5	4
0.25	38,266.61	38,265.55	38,278.28	38,286.31	38,280.58	38,288.88	38,301.41	38,315.68
0.50	38,267.06	38,268.54	38,270.15	38,277.30	38,282.53	38,288.59	38,301.42	38,296.08
0.75	38,267.05	38,269.10	38,271.47	38,275.93	38,280.49	38,286.52	38,296.44	38,307.47
1.00	38,267.06	38,268.43	38,271.78	38,275.14	38,280.05	38,283.93	38,292.89	38,301.26
1.25	38,266.99	38,268.33	38,271.79	38,276.54	38,281.75	38,289.31	38,297.22	38,298.64
1.50	38,266.90	38,268.21	38,270.69	38,273.16	38,282.96	38,288.87	38,291.20	38,302.78
1.75	38,266.96	38,268.75	38,270.82	38,277.40	38,279.58	38,286.46	38,289.78	38,305.46
2.00	38,266.90	38,268.56	38,270.90	38,278.06	38,282.07	38,286.52	38,286.83	38,298.37

	$\boldsymbol{\sigma}$							
$\boldsymbol{h}$	0.5	1	1.5	2	2.5	3	3.5	4
0.25	1.21	2.69	4.58	6.26	6.50	8.06	10.10	12.56
0.50	0.13	0.31	0.60	1.48	1.66	3.24	4.40	3.81
0.75	0.09	0.40	0.64	1.15	2.08	2.08	4.46	5.84
1.00	0.10	0.27	0.66	1.16	1.93	2.42	3.07	4.67
1.25	0.07	0.24	0.72	1.36	1.68	3.09	3.95	4.77
1.50	0.06	0.27	0.57	0.83	1.96	2.55	3.52	4.99
1.75	0.06	0.30	0.55	1.60	1.78	2.19	2.96	5.62
2.00	0.05	0.25	0.56	1.76	2.41	2.87	2.41	4.25

Table 7.2: Loss Funtion. First table shows the value of $L(\boldsymbol{x}) = (1/n)\sum_{i=1}^{n} L(x_i)$ for the heteroscedastic model with a mean function $f_1(x) = \sin^3(2\pi x^3)$ and error from $\mathcal{N}(0, 16)$. Second table shows standard errors for the corresponding average values of the loss function given in the table above.

and it becomes smaller and levels out with increased h. Table 7.3 shows the results of the smooth exponential function f_2, from which we notice that the average value of the loss function $L(\boldsymbol{x})$ is affected not just by the change in σ^2, but also by the change in value of the chosen bandwidth parameter h. The value of $L(\boldsymbol{x})$ function increases for a given value of h with increasing σ^2. For a given value of σ^2, the value of the $L(\boldsymbol{x})$ is the largest for the smallest value of h, then changes with no discernible pattern over increases in h.

When the leave-one-out CV bandwidth parameter choice is used the variability in the estimate becomes apparent, as this become evident in the case of the sinusoid function $f_2(x)$. In the second part of the study in which we had controlled the value of the bandwidth parameter the changes in the loss function are fairly small (Table 7.2),

	σ							
h	0.5	1	1.5	2	2.5	3	3.5	4
0.25	29.21	71.52	147.60	211.88	347.87	623.79	790.56	779.92
0.50	19.42	35.41	69.92	97.14	128.00	212.04	265.10	347.41
0.75	17.51	34.32	50.82	76.13	171.06	176.89	251.48	352.97
1.00	19.12	32.93	66.29	76.24	129.98	180.38	168.74	341.20
1.25	22.00	28.53	53.57	84.80	141.25	212.80	260.93	296.38
1.50	19.52	39.90	52.88	89.65	139.09	199.33	324.59	236.20
1.75	17.48	34.97	53.00	90.84	160.73	166.97	260.61	334.66
2.00	17.93	29.83	53.15	118.05	135.23	178.88	215.32	306.09

	σ							
h	0.5	1	1.5	2	2.5	3	3.5	4
0.25	2.81	11.40	18.77	21.97	41.91	78.31	92.89	96.51
0.50	1.54	3.75	7.13	14.23	16.56	24.20	35.48	46.48
0.75	1.21	3.05	5.96	9.76	24.91	19.18	29.22	41.54
1.00	1.43	3.33	8.64	8.44	16.13	23.87	22.44	42.97
1.25	1.79	3.00	6.01	9.52	16.06	33.37	37.11	40.39
1.50	1.16	4.19	5.22	11.08	16.92	25.10	42.30	38.79
1.75	1.14	3.08	5.80	10.93	20.19	22.02	37.09	54.46
2.00	1.08	2.60	6.14	13.62	15.06	22.30	28.34	34.29

Table 7.3: Loss Funtion. First table shows the value of $L(\boldsymbol{x}) = (1/n)\sum_{i=1}^{n} L(x_i)$ for the heteroscedastic model with a mean function $f_2(x) = (\frac{1}{4}\exp(2x))$ and error from $\mathcal{N}(0, 16)$. Second table shows standard errors for the corresponding average values of the loss function given in the table above.

but when leave-one-out CV bandwidth parameter is used for the same heteroscedastic case, the variability in the loss function becomes more obvious (Table 7.1).

In the difference-based approach for the selection of the bandwidth parameter in the local polynomial fit, we use the K-fold cross-validation for correlated data method. Levine (2006) in his paper presents a case study in which he observes the influence of the choice of parameter K : $K = 5, 10$ and 15. By observing the cross-validation discrete mean square error $CDMSE = n^{-1}\sum_{i=1}^{n}\left(\hat{f}_{h_{CV}}(x_i) - f(x_i)\right)^2$ with h_{CV} defined by (7.18) and for any given smooth function $f(x)$, he comes to the conclusion that for the relatively small sample size of $n = 500$, increasing K reduces the variability of CDMSE value. He suggests that in general, the prevalent choice should be $K = 5$ or 10, although he regards the choice of $K = 10$ as "probably" the

better option. We have run our own test in which we compare the performance of the difference-based approach for the different values of the parameter K. We again define the squared error loss function, this time as the mean of $L(x_i) = \left(\hat{f}(x_i) - f(x_i)\right)^2$, where $f(x) = (\frac{1}{4}\exp(2x))$. As the most difficult case, we consider a heteroscedastic problem with the error from the normal distribution, $e_i \sim \mathcal{N}(0, 16)$. The design is equally spaced $x_i = 1/n$, $i = 1, \ldots, n$, for $n = 500$. The number of the realisation in the study is 100 and the choices for K are 5 and 10. The simulation study for values of $K = 5$ and 10 shows a slightly smaller $L(\boldsymbol{x})$ for $K = 10$ (See Table 7.4). Considering the computational time required and this result we chose $K = 10$ as a more preferable for our study in Example 5.

average value of the loss function		
	$K = 5$	$K = 10$
$L(\boldsymbol{x})$	0.54210	0.54184

Table 7.4: Loss Function. Average values of the squared error loss function, $L(x) = \left(\hat{f}(x_i) - f(x_i)\right)^2$, where $f(x) = (\frac{1}{4}\exp(2x))$ when applying difference-based approach using local polynomial of order one for which the bandwidth parameter h is selected by applying K-fold cross-validation for correlated data, with $K = 5$ and 10.

The difference-based approach requires the choice of the order of the difference. In our simulation study we use the first order difference. We do not use higher order estimators due to the reason given earlier. The following list indicates the variance function estimation methods used in this study:

1) double generalised linear model from Tweedie family with the log link *dglm*;

2) residual-bassed variance estimation using local polynomial of order one *LocPol-p1-res*;

3) difference-based variance estimation using local polynomial of order one *LocPol-p1-diff*;

4) difference-based variance estimation using local polynomial of order three *LocPol-p3-diff*.

Figure 7.3 illustrates fitted conditional variance functions for four homoscedastic models with the error from four different distributions: $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$. Figure 7.4 and Figure 7.5 show the fitted variance functions for heteroscedastic models with the error from four different distributions: $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$ given in Example 5, when the variance function is $\sigma^2(x_i) = 1 + 4x_i$ and $\sigma^2(x_i) = 0.25\exp(2x_i)$, respectively. The full list of the results for this simulation study is given in the Appendix E.

7.4.1 Model 1: Homoscedastic, Location-Shift Model

All four methods used for the estimation of the conditional variance function, when put in the context of our bootstrapping approach provide very good results. The model parameter estimates of β_0 and β_1 are very close to their true values. In the models with errors from $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$ and $t(20)$ the intercept parameter is slightly over-estimated and the slope parameter slightly under-estimated. For the model with the error from $t(10)$ the intercept has the same properties as the other models, but the slope parameter estimate is method dependent. It is slightly under-estimated when using parametric approach for the conditional variance estimation (dglm) and this is the same for the rest of the competing bootstrapping techniques used with the exception of the MCMB. When non-parametric methods for the estimation of the conditional variance are used the slope parameter is slightly over-estimated, just as when applying the rank-score and Wald methods. The coverage probability for the estimated parameters for all four models is high and very similar for all of the approaches used. The exception is the rank-score method that although providing the shortest lengths of the parameters' confidence intervals the coverage probabilities are the poorest making this approach less desirable than the others as it has been already reported in the previous studies. Considering this, the following analysis of the results exclude the rank-score as a fair competitor.

The average lengths of the confidence intervals for b_0 and b_1 from models with error from the normal distribution are the shortest when applying difference-based

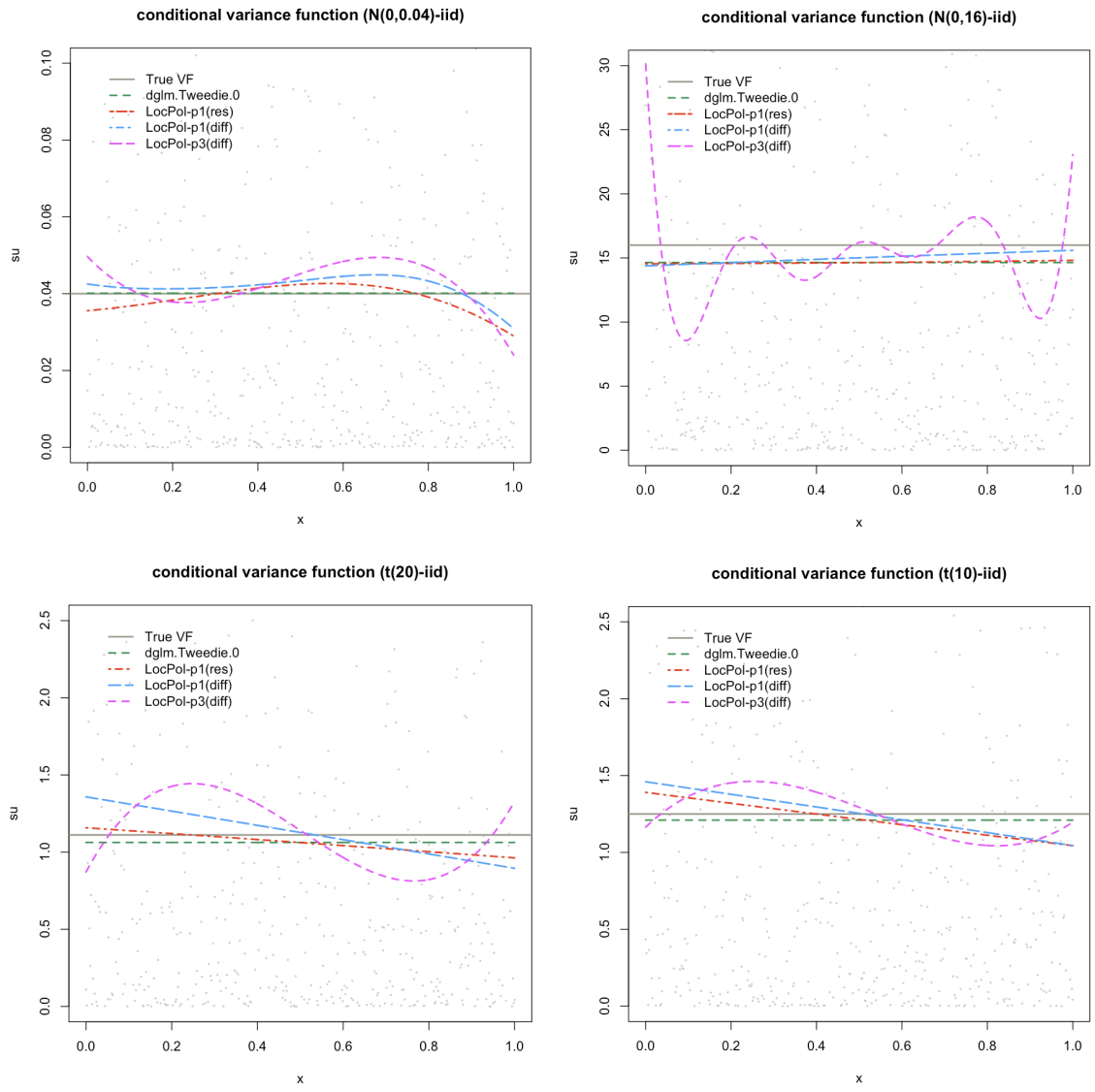


Figure 7.3: *Conditional Variance Function Estimations for Example 5 homoscedastic models (M1). The figures plot the squared centered residuals of the simulated data for the homoscedastic models in Example 5 with the errors from four different distributions:* $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ *and* $t(10)$. *Solid gray line on the plots represents the true variance function. The estimates of the conditional variance functions are given in different colored dashed lines: green, red, turquoise and magenta, when using the following methods: dglm, LocPol-p1-res, LocPol-p1-diff, LocPol-p3-diff, respectively.*

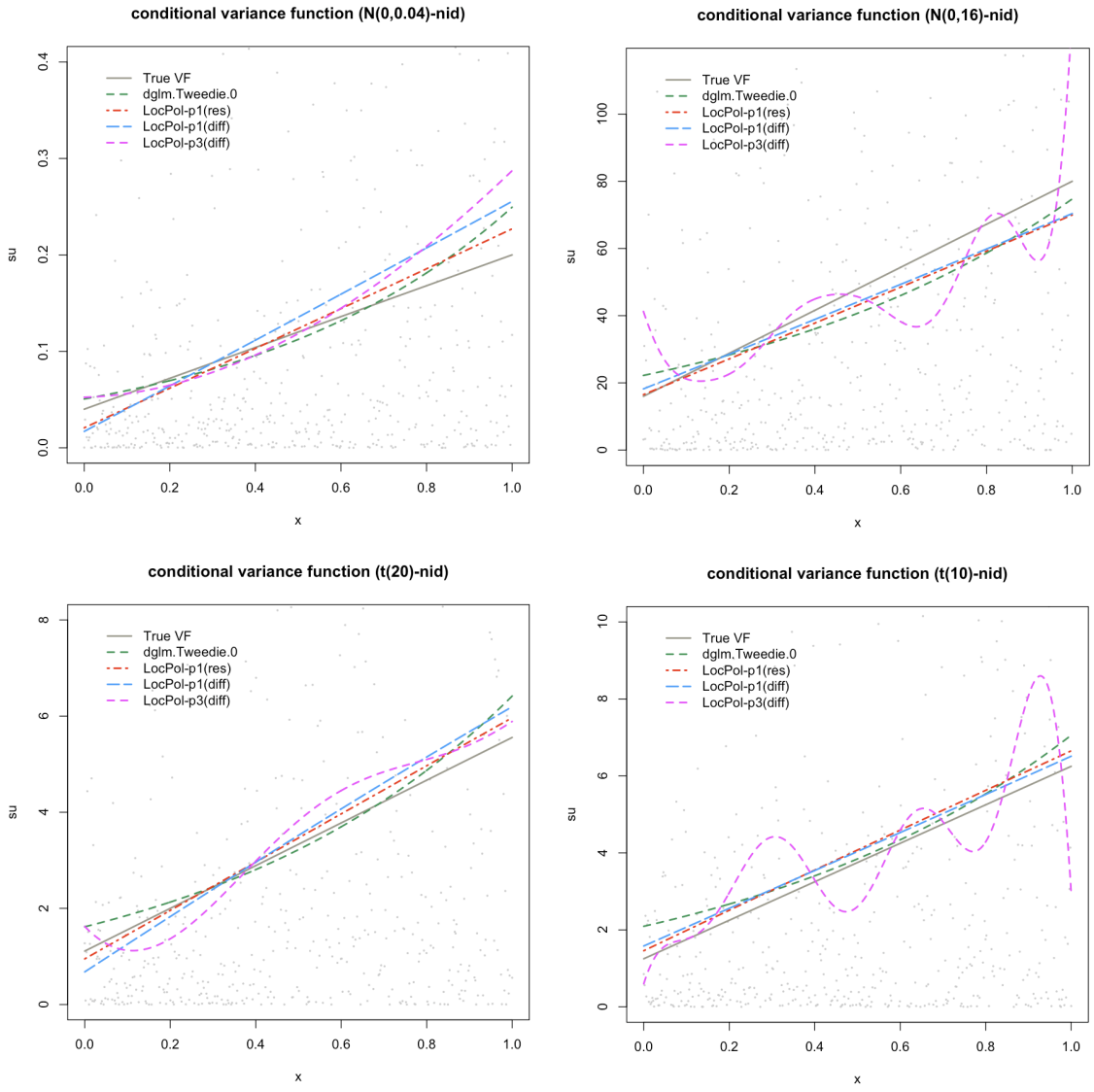


Figure 7.4: *Conditional Variance Function Estimations for Example 5 heteroscedastic models (M2). The figures plot the squared centered residuals of the simulated data for the heteroscedastic models in Example 5 with the errors from four different distributions:* $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ *and* $t(10)$. *Solid gray line on the plots represents the true variance function. The estimates of the conditional variance functions are given in different colored dashed lines: green, red, turquoise and magenta, when using the following methods: dglm, LocPol-p1-res, LocPol-p1-diff, LocPol-p3-diff, respectively.*

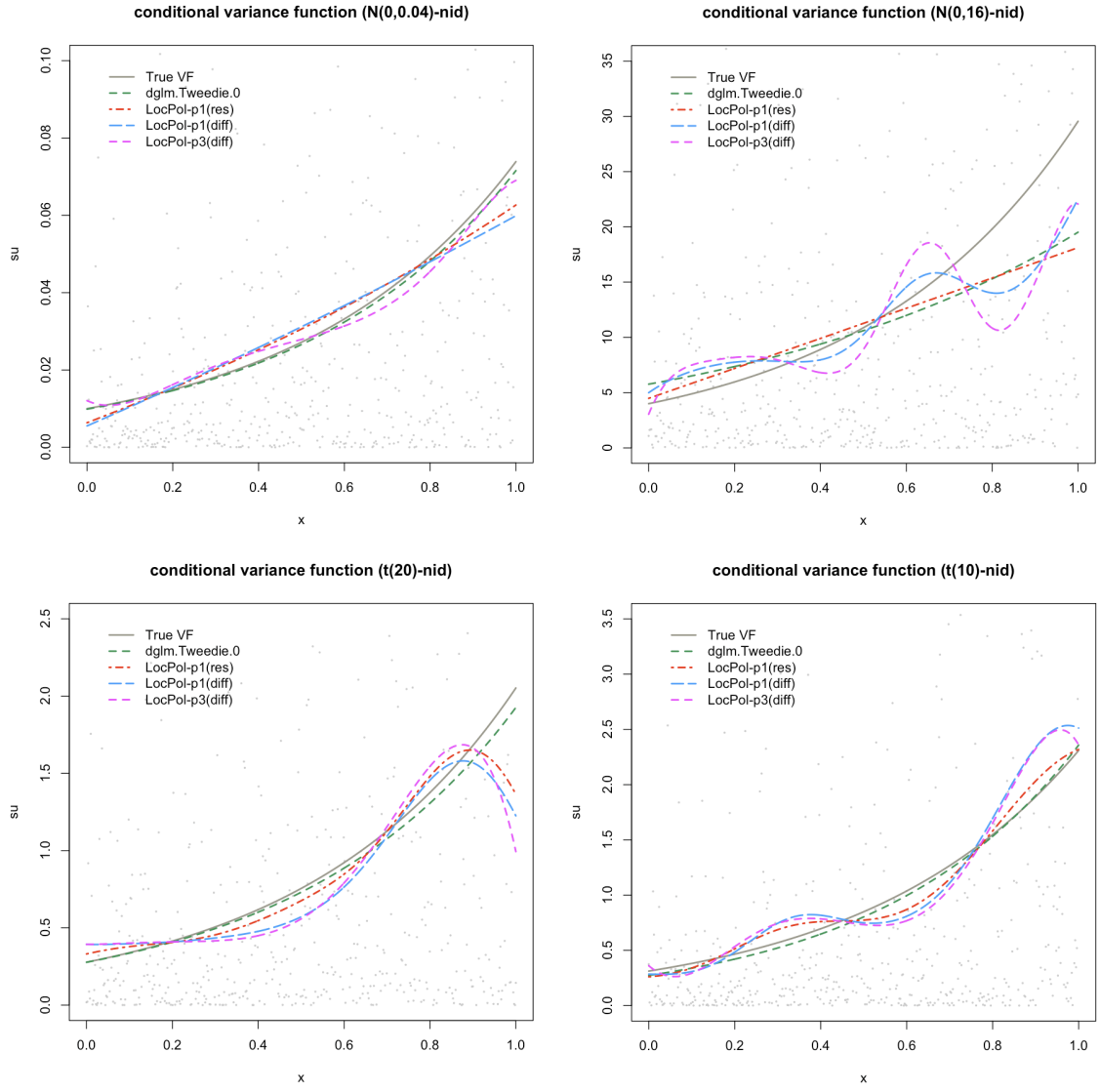


Figure 7.5: *Conditional Variance Function Estimations for Example 5 heteroscedastic models (M2). The figures plot the squared centered residuals of the simulated data for the heteroscedastic models in Example 5 with the errors from four different distributions:* $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ *and* $t(10)$. *Solid gray line on the plots represents the true variance function given by* $\sigma^2(x_i) = 0.25 \exp(2x_i)$. *The estimates of the conditional variance functions are given in different colored dashed lines: green, red, turquoise and magenta, when using the following methods: dglm, LocPol-p1-res, LocPol-p1-diff, LocPol-p3-diff, respectively.*

approach using local polynomial fit of order one. For the other two models with error from t distribution double GLM with Tweedie family with *log* link provides the best results, and if the t distribution has a very long tail non-parametric local polynomial fit to the squared residuals can be considered as a good alternative approach.

Using the difference-based approach with the local polynomial of order three in the kernel smoothing bootstrapping does not provide good results, this method provides the widest confidence intervals for the model parameters. The average values of the variance of the standardised residuals are slightly over-inflated and the general variability of the parameter estimates is high compared to the other methods. Considering these results, this method appears to be the least desirable for the estimation of the conditional variance function here.

7.4.2 Model 2: Heteroscedastic, Location-Scale Model

In this study we have consider two different variance functions:

a) $\sigma^2(x_i) = 1 + 4x_i$ and

b) $\sigma^2(x_i) = \frac{1}{4}\exp(2x_i)$.

For all of the heteroscedastic models considered in this study we get estimates of the model parameters β_0 and β_1 which are very close to the their true values. In each model, the parameters are either consistently slightly over-estimated or slightly-under estimated by all of the approaches used. When the error in the model comes from the $t(10)$ distribution the parameter estimates are more inconsistent. All of the approaches used provide high and very similar coverage probabilities. Again, the exception being the rank-score method that provides the shortest lengths of the parameters' confidence intervals, but the lowest coverage probabilities. Taking this into consideration, as in the analysis of the results for homoscedastic models, the following examination of the obtained results excludes the rank-score method.

In the models with the linear variance function (a), competitive results are consistently obtained by the kernel smoothing bootstrapping method in which the conditional variance function is estimated by the double general linear model with Tweedie

family with log link. When observing the results obtained by the kernel smoothing bootstrapping, in the model with the normal error with high variance ($\sigma^2 = 16$) the shortest average lengths of the confidence intervals for b_0 and b_1 parameters are obtained when the conditional variance is estimated using DGLM. In the model with $\sigma^2 = 0.04$ that is the case when the difference-based approach with the local polynomial fit of order one is used. For the other two models with the errors from the t distributions, the shortest average lengths of the confidence intervals for b_0 and b_1 parameters are obtained by estimating the conditional variance using a local polynomial of order one fitted to the squared residuals.

For the models with non-linear variance function (b) the kernel smoothing bootstrapping approach provide the shortest average lengths of the confidence intervals for the b_0 and b_1 parameters, using DGLM for the estimation of the conditional variance. The only exception is the model with the normal error with variance $\sigma^2 = 0.04$, for which a local polynomial fit to the squared residuals provides to be a better choice.

Observing the average values of the ratio of the determinants of the models with the linear variance function we notice high variability in the estimation of the parameters when using difference based approaches, in particular when fitting a local polynomial of order three. Observing the same for the models with the non-linear variance function all of the non-parametric methods used for the estimation of the conditional variance function show high variability in their estimation. We still have slightly wider mean length of the confidence interval of the intercept parameter when using our kernel smoothing bootstrapping technique in comparisons with the other bootstrapping methods used. However, we have to emphasize that their standard errors are in some cases more than the half the size of the competing bootstrapping methods (*xy*, *pwy*, *mcmb* and *wxy*). The exceptions are kernel bootstrapping methods in which the conditional variance function has been estimated using the difference based approach. For those approaches we also notice the highest standard errors of the average lengths of the parameters confidence intervals.

7.5 Conclusion

Analysing the results given in the Appendix E we notice that the difference-based approach offers very little improvement to our kernel smoothing bootstrapping approach. The benefits are noticeable only in the homoscedastic models with normal errors and the most simple heteroscedastic model with the linear variance function and the error from $\mathcal{N}(0, 0.04)$. The advantage in using the difference-based approach is provided only if the first order local polynomial is fitted to the calculated differences. Applying a higher order local polynomial to the calculated differences appears to yield high variability of the model estimation in three ways. This is reflected by:

i. significantly higher standard errors of the average lengths of the parameter confidence intervals;

ii. over inflated average values of the variance of the standardsied residuals;

iii. extremely high average value of the ratio of the determinants in some cases.

In general, it appears that for the heteroscedastic models and in particular those with the non-linear variance function, the variability in the estimates increases when using the difference-based approach, making it one of our least desirable choices for modeling the conditional variance function.

In the most complex homoscedastic models in which the error is normally distributed with high variance ($\sigma^2 = 16$) or non-normally distributed, estimating the conditional variance function using double general linear model with Tweedie family with *log* link has provided the best results. This is the case for the majority of other simpler models, when the conditional variance is estimated by fitting the local polynomial of order one to the squared residuals.

For homoscedastic models kernel smoothing bootstrapping method provides very good results, regardless of the modeling approach used for estimation of the conditional variance function. Comparing it with the other competing bootstrapping techniques, the kernel smoothing bootstrapping method slightly over-inflates the mean length of the confidence interval of the intercept parameter in the heteroscedastic

models. However, the standard errors of the average lengths of the confidence intervals for both coefficients are much smaller in comparisons to the others. This makes our bootstrapping approach less variable in its estimation.

Despite its popularity, the difference-based approach for modeling the conditional variance still has a lot of questions that remain unanswered and that require further investigation, especially from the practical point of view. From our case study it appears that the method is sensitive to the non-normal distribution and non-constant variance. In particular, the choice of the bandwidth parameter used has proven to be difficult as the application of the popular cross-validation approach is jeopardized by the correlated data. In order to remove the correlation of the k-dependent sequence from the k^{th} order differences we have proposed the $1 + 2k$ cross-validation method.

We run another test to compare the performance of the first order difference-based approach in which the bandwidth parameter used is selected by K-fold CV ($K = 10$) method and $1 + 2k$ CV method. Going back to the previously defined squared error loss function, $L(x) = \left(\hat{f}(x) - f(x)\right)^2$, where $f(x) = (\frac{1}{4}\exp(2x))$ we consider a heteroscedastic problem with the error from the normal distribution, $e \sim \mathcal{N}(0, 16)$. The design is equally spaced $x_i = 1/n$, $i = 1, \ldots, n$, for $n = 500$ and we run 100 realisations. By using the difference-based approach and applying two different bandwidth parameter selection methods, Table 7.5 shows there is a small difference in estimation efficiency achieved.

average value of the loss function		
	K-fold CV	$1 + 2k$ CV
$L(\boldsymbol{x})$	0.50237	0.50400

Table 7.5: Average values of the squared error loss function, $L(x) = \left(\hat{f}(x_i) - f(x_i)\right)^2$, where $f(x) = (\frac{1}{4}\exp(2x))$ when applying difference-based approach using the bandwidth parameter for the local polynomial of order one fitting when selected by K-fold CV and $1 + 2k$ CV methods, where K=10 and k=1.

In order to assess the performance of our proposed $1+2k$ bandwidth selection technique for the difference based approach when used in the context of kernel smoothing

bootstrapping we run another small simulation study. For the estimation of the conditional variance function in the kernel smoothing bootstrapping approach we use the following methods:

i) Tweedie DGLM with link=*log*;

ii) Joint Modelling (using Tweedie GLM link=*log* for the modeling of the dispersion);

iii) Local Polinomial of order one (residual-based approach);

iv) Local Polinomial of order one (differencel-based approach) using K-fold CV method for the bandwidth selection ($K = 10$);

v) Local Polinomial of order one (differencel-based approach) using $1 + 2k$ CV for the bandwidth selection.

We use the most complex heteroscedastic model (M2) with the non-linear variance function $\sigma^2(x) = \frac{1}{4}exp(2x)$, and the error term from the four different distributions: $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$. The design is equally spaced $x \in [0, 1]$ where $x_i = i/n$ and with $n = 500$. Considering that the $1 + 2k$ CV is time consuming, for each model we run 30 realisations ($R = 30$) and we use a thousand bootstraps ($B = 1,000$) for the estimation of the 50^{th} quantile. As the $1 + 2k$ CV for increased k becomes wasteful, we use the first-order differences ($k = 1$) for the difference based approaches. To make an informative comparison, we judge the performance of the used methods by the average lengths (L) and coverage probability (C) of the computed 95% confidence intervals (CI) for each of the coefficients in the model. The variability of the model parameter estimates is examined by the ratio of the determinants of the estimated covariance matrixes.

As in the previous studies, all of the applied methods provide close estimation of the model parameters β_0 and β_1 (see Table 7.6). For the models with the error from $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$ and $t(20)$ the intercept parameter is slightly under-estimated and the slope parameter slightly over-estimated. In the model with the error from

$t(10)$, we get the opposite effect. The coverage probabilities are very high and similar for all of the approaches used (see Table 7.7).

As mentioned in the previous study, for the heteroscedastic models with the non-linear variance function our preferred choice for the estimation of the conditional variance is DGLM. It can be seen from Table 7.7 that using this method for the estimation of the conditional variance provides the shortest average lengths of the confidence intervals with very small standard errors. Comparing the outputs obtained when using difference-based approaches we notice that as long as we do not depart greatly from the normal distribution the choice of the bandwidth parameter h, that is used in the local polynomial fitting, appears to be better when $1+2k$ CV method is used. In the models with the error from : $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$ and $t(20)$ the average length of the confidence intervals is shorter when using $1+2k$ CV rather than $K = 10$-fold CV method for the selection of the parameter h. In these three models, when using a difference-based approach with the bandwidth parameter selected by K-fold CV, the variability of the estimated parameters b_0 and b_1 is very high. This is reflected through the much wider lengths of the CI's that have the highest standard errors, over-inflated average value of the variance of the standardised residual (see Table 7.8) and high average value of the ratio of the determinants of the estimated covariance matrixes (see Table 7.9). However, in the model with the error from the long tail distribution, $t(10)$, when the choice of the bandwidth parameter is made by using the K-fold CV the difference-based approach provides better results than when using the $1 + 2k$ CV method for the selection of h. For this model with error from $t(10)$ by demonstrating high variability, the worst performing approaches are difference-based approach with the $1 + 2k$ CV selection of the bandwidth parameter and the residual-based approach using the local polynomial of order one.

From the analysis of the results obtained in this study we conclude that the choice of the method used for the estimation of the conditional variance function depends not only on the type of the regression model we deal with, but also on the underlying variance function itself. We have seen that for the majority of homoscedastic models analysed in this study estimating the conditional variance function using double

general linear model with Tweedie family with *log* link provides the shortest lengths of the parameters' CI's with small standard errors (regarded as best results). In the heteroscedastic cases in which the variance function is linear we get best results by applying a local polynomial of order one fitted to the squared residuals for its estimation. However, for those kind of models, these results are closely followed by the parametric approaches used for the estimation of the conditional variance. The best results for the heteroscedastic models with a non-linear variance function are obtained when using DGLM with Tweedie family with *log* link for their estimation. Considering these findings and the complexity of real, practical problems we suggest the parametric estimation of the conditional variance function using DGLM with Tweedie family with *log* link to be applied to the standardisation of the residuals used in the kernel smoothing bootstrapping adjusted to have the same mean and the variance as the data from which it is constructed.

In the last two studies in which we focus on comparing the performances of the difference-based approache when selecting the bandwidth parameter by the K-fold and the $1 + 2k$ CV methods we have noticed a high variability in the estimation of the bandwidth parameter h. Considering that when using our kernel smoothing bootstrapping approach we get slightly over-inflated CI's for the intercept parameter in the heteroscedastic models we have tried to adopt $1 + 2k$ CV to accommodate heteroscedasticity. We have considered the idea proposed by Ruppert et al. (1995) of segmenting the data into the blocks and estimating the bandwidth parameter for each section of the data. In the heteroscedastic models, which we consider in our studies, it is sensible to expect a different degree of smoothing to be required for each block. As the variability increases with x, we expect higher values for the estimated h as we move through the data blocks. However, this was not the case. Regardless of the changes introduced to the model by altering the variability of the error term and changing the number of the blocks used for the segmentation of the data, the estimates of h from one block to the other were very unbalanced. We have tried to introduce the robust approaches to minimise the influence of the extreme values in each of the blocks, by introducing different weighting schemes, but unfortunately we

	β_0		β_1	
Method	b_0	SE	b_1	SE
$e \sim \mathcal{N}(0, 0.04)$				
dglm.Tw.0	1.99750	0.00212	5.01074	0.00450
JointMod	1.99757	0.00209	5.01101	0.00448
LocPoly-ress	1.99738	0.00208	5.01125	0.00445
LocPoly-diff-K10-fold	1.99737	0.00212	5.01065	0.00446
LocPoly-diff-1+2k	1.99771	0.00213	5.01056	0.00454
$e \sim \mathcal{N}(0, 16)$				
dglm.Tw.0	1.95009	0.04242	5.21488	0.08990
JointMod	1.95134	0.04184	5.22021	0.08950
LocPoly-ress	1.94759	0.04162	5.22504	0.08895
LocPoly-diff-K10-fold	1.96573	0.04530	5.18378	0.09144
LocPoly-diff-1+2k	1.95423	0.04263	5.21108	0.09088
$e \sim t(20)$				
dglm.Tw.0	1.99756	0.01243	5.03616	0.02984
JointMod	1.99762	0.01240	5.03651	0.02966
LocPoly-ress	1.99813	0.01246	5.03496	0.02972
LocPoly-diff-K10-fold	1.99870	0.01242	5.03721	0.03038
LocPoly-diff-1+2k	1.99835	0.01244	5.03504	0.02974
$e \sim t(10)$				
dglm.Tw.0	2.00466	0.01165	4.98942	0.02852
JointMod	2.00376	0.01167	4.99079	0.02874
LocPoly-ress	2.00548	0.01181	4.99058	0.02878
LocPoly-diff-K10-fold	2.00482	0.01169	4.98813	0.02850
LocPoly-diff-1+2k	2.00247	0.01197	4.98982	0.02838

Table 7.6: Parameter estimates of the 50^{th} quantile for the heteroscedastic Model 2: $y_i = 2 + 5x_i + \sigma(x_i)e_i$ with the variance function: $\sigma^2(x) = \frac{1}{4}exp(2x)$.

	β_0			β_1		
Method	C	L	SE	C	L	SE
$e \sim \mathcal{N}(0, 0.04)$						
dglm.Tw.0	100.00%	0.06863	0.00058	100.00%	0.11947	0.00099
JointMod	100.00%	0.06949	0.00052	100.00%	0.12114	0.00100
LocPoly-ress	100.00%	0.06932	0.00057	100.00%	0.12013	0.00104
LocPoly-diff-K10-fold	100.00%	0.07231	0.00261	100.00%	0.12509	0.00462
LocPoly-diff-1+2k	100.00%	0.06969	0.00049	100.00%	0.12044	0.00095
$e \sim \mathcal{N}(0, 16)$						
dglm.Tw.0	100.00%	1.37248	0.01170	100.00%	2.38946	0.01979
JointMod	100.00%	1.38982	0.01033	100.00%	2.42289	0.01996
LocPoly-ress	100.00%	1.38653	0.01144	100.00%	2.40264	0.02084
LocPoly-diff-K10-fold	100.00%	3.24687	1.30284	100.00%	5.71045	2.32127
LocPoly-diff-1+2k	100.00%	1.39384	0.00987	100.00%	2.40883	0.01894
$e \sim t(20)$						
dglm.Tw.0	100.00%	0.35325	0.00390	93.33%	0.61410	0.00681
JointMod	100.00%	0.35212	0.00368	93.33%	0.61116	0.00693
LocPoly-ress	100.00%	0.35203	0.00395	96.67%	0.60752	0.00703
LocPoly-diff-K10-fold	100.00%	0.40961	0.05596	93.33%	0.71031	0.09822
LocPoly-diff-1+2k	100.00%	0.35389	0.00312	93.33%	0.61324	0.00554
$e \sim t(10)$						
dglm.Tw.0	100.00%	0.36120	0.00382	96.67%	0.62200	0.00634
JointMod	100.00%	0.36103	0.00432	96.67%	0.62605	0.00733
LocPoly-ress	100.00%	0.41797	0.05774	93.33%	0.72373	0.10057
LocPoly-diff-K10-fold	100.00%	0.36249	0.00353	93.33%	0.62571	0.00697
LocPoly-diff-1+2k	100.00%	0.47024	0.07775	96.67%	0.81774	0.13447

Table 7.7: Results for the parameter estimates of the 50^{th} quantile for the heteroscedastic Model 2: $y_i = 2 + 5x_i + \sigma(x_i)e_i$ with the variance function: $\sigma^2(x) = \frac{1}{4}exp(2x)$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean of the standardised residuals	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	1.00806	0.00464
JointMod	1.02338	0.00210
LocPoly-ress	0.99665	0.00362
LocPoly-diff-K10-fold	1.30762	0.20515
LocPoly-diff-1+2k	1.00044	0.00312
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	1.00806	0.00463
JointMod	1.02338	0.00210
LocPoly-ress	0.99665	0.00362
LocPoly-diff-K10-fold	7.88508	4.84493
LocPoly-diff-1+2k	1.00044	0.00312
$e \sim t(20)$		
dglm.Tw.0	1.01737	0.00981
JointMod	1.02273	0.00273
LocPoly-ress	1.00156	0.00486
LocPoly-diff-K10-fold	2.31136	1.27230
LocPoly-diff-1+2k	1.00951	0.00584
$e \sim t(10)$		
dglm.Tw.0	1.01026	0.00563
JointMod	1.02374	0.00366
LocPoly-ress	1.79033	0.79099
LocPoly-diff-K10-fold	1.01232	0.00770
LocPoly-diff-1+2k	2.20944	0.83366

Table 7.8: Results for the parameter estimates of the 50^{th} quantile for the heteroscedastic Model 2: $y_i = 2 + 5x_i + \sigma(x_i)e_i$ with the variance function: $\sigma^2(x) = \frac{1}{4}exp(2x)$. Mean of the standardised residuals when using Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach with bandwidth h selected by K-fold CV method and Local Polynomial of $p = 1$ difference-based approach with bandwidth h selected by $1 + 2k$ CV method.

Method	mean of the ratio of the determinants	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	0.93462	0.03606
JointMod	0.97433	0.03860
LocPoly-ress	0.96027	0.03493
LocPoly-diff-K10-fold	1.27169	0.22904
LocPoly-diff-1+2k	0.95754	0.03420
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	0.93462	0.03607
JointMod	0.97437	0.03856
LocPoly-ress	0.96019	0.03486
LocPoly-diff-K10-fold	7094.44953	4928.87395
LocPoly-diff-1+2k	0.95763	0.03420
$e \sim t(20)$		
dglm.Tw.0	0.98622	0.03831
JointMod	0.96780	0.04174
LocPoly-ress	0.96749	0.03425
LocPoly-diff-K10-fold	59.90431	58.93610
LocPoly-diff-1+2k	0.96480	0.03338
$e \sim t(10)$		
dglm.Tw.0	1.07137	0.04746
JointMod	1.07820	0.04989
LocPoly-ress	35.78808	34.71722
LocPoly-diff-K10-fold	1.07903	0.04496
LocPoly-diff-1+2k	61.14498	42.76797

Table 7.9: Results for the parameter estimates of the 50^{th} quantile for the heteroscedastic Model 2: $y_i = 2 + 5x_i + \sigma(x_i)e_i$ with the variance function: $\sigma^2(x) = \frac{1}{4}exp(2x)$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed. Five different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach with bandwidth h selected by K-fold CV method and Local Polynomial of $p = 1$ difference-based approach with bandwidth h selected by $1 + 2k$ CV method.

did not manage to get a desirable outcome. It appeared that the CV approach is not just sensitive to the outliers but also to the skewness in the data. Loader (1999) reports the same difficulties even when observing only homoscedastic models. He reports that even when the samples are drawn from the same model, cross validation selects bandwidths that are very different from sample to sample and the presence of the heavy tails makes the estimate of the bandwidth parameter much more difficult. This becomes even more of an issue in the heteroscedastic cases and specially for our problem in which we deal with the squared differences and squared residuals that creates heavy long tails in the error distribution of the regression model, which is challenging.

Chapter 8

Construction of the Reference Growth Charts Using the LMS Method

8.1 Introduction

Quantile regression has a wide application in medial science for the construction of the growth standards. Growth charts display the distribution of a certain physical measurement within a certain range of time for a certain population. They are used to screen the measurement form an individual subject in the context of population values. They also can be used to track an individual's measurement over time.

The construction of the growth curves requires the flexibility that is not necessarily offered by a parametric approach which makes rigid assumptions about the functional form of the fitted quantiles. In this chapter we discuss various issues involved in using the LMS (Lamda-Mu-Sigma) method of Cole and Green (1992) for constructing the reference growth standards that offers the flexibility in smoothing the growth curves. The LMS method provides a way of obtaining growth standards for healthy individuals and is based on normalizing the conditional distribution of a measure using the power transformation of Box and Cox (1964). The package `lmsqreg` developed by Carey (2002) implements the LMS method in `R`.

The application of the LMS method is illustrated using the Saudi Arabian data, which was collected as a part of a nationwide project to establish normal anthropometric measurements for Saudi Arabian children and adolescents. The growth standards are derived from a cross sectional sample of healthy children and adolescents aged from birth to 19 years. The sample was randomly selected by a stratified multistage probability sampling procedure from each of the 13 administrative regions of the Kingdom of Saudi Arabia, ensuring both national and urban/rural representation. The anthropometric data comprises 51,485 observations of which 25,987 are made on boys and 25,498 on girls. Those measurements include: length, for the children 2 years of age and below, height, for children above 2 years of age, weight and head circumference.

The reference growth charts we have constructed describe the dependance of height, weight, body mass index (BMI) and head circumference on age, and weight on length/stature for two age ranges, birth to 36 months and 2 to 19 years. Use of the LMS method was a requirement of the study. Issues related to the application of the LMS methodology are specifically illustrated using the Saudi girls weight data for those from birth to three years old.

8.2 Outliers

An *Outlier* is a sample value that lies outside the main pattern or distribution of the data and in the context of quantile regression, which was first introduced by Koenker and Bassett (1978), it will be one which has a much larger or smaller response value at a given age when compared with other responses at a similar age. Quantile regression measures the effect of covariates not only in the center of the distribution but also in the upper and lower tails. Extremely low and extremely upper quantiles are of interest regarding growth charts and therefore it is important to deal with the issue of removing the potential outliers with a cautiousness.

The outlier should not be regarded as a pejorative term; outliers may be correct, but they should be checked for transcription error (Venables and Ripley, 2002). The

quantile regression model is a natural extension of the linear regression model. If an outlier is included in the data which is used to estimate the quantiles then it may be highly influential on the fitted regression line in that the line may be pulled in a disproportionate manner towards the outlying value or it may cause a failure in the algorithm used to estimate the quantiles (Koenker and Bassett, 1978). This latter point is particularly true with respect to the LMS procedure, as according to Carroll (1982) the choice of the transformation L(x) is highly sensitive to outliers in the data. We have also found that if the outliers are not removed it can result in the numerical failure of the model fitting algorithm in the function `lmsqreg`.

The lack of a methodology to assess the direct effect of an individual observation on the LMS methodology has prompted us to approximate the LMS model using a cubic regression line to model the relationship between a response and covariate (such as weight and age). Approximating the LMS model in this way enables us to identify the outliers in that space with respect to this mode, that hopefully are also the outliers with respect to the LMS model. To fit this cubic regression line we have used a robust regression procedure.

Robust regression deals with cases that have very high leverage, and cases that are outliers. Robust regression represents a compromise between the efficiency of the ordinary least squares (OLS) estimators and the resistance of the least absolute value (LAV) estimators, both of which can be seen as special cases of M-estimation (Huber, 1981). It is a form of weighted least squares regression, which is similar to least square in that it uses the same minimization of the sum of the squared residuals, but it is done iteratively. Based on the residuals a new set of weights are determined at each step. In general, the larger the residuals, the smaller the weights. So the weights depend on the residuals. At the same time, the residuals depend on the model and the model depends on the weights. This generates an iterative process and it goes on until the change in the parameter estimates are below a preset threshold. At the end, instead of all points being weighted equally, the weights vary and those with the largest weights contribute more to the fit.

There are a few types of weighting schemes, M-estimetors, that can be implemented (Venables and Ripley, 2002). In Huber's (1981) weighting, observations with small residuals get a weight of 1, the larger the residual, the smaller the weight. M-estimetion, introduced by Huber (1964) can be regarded as a generalisation of maximum-likeliood estimation (MLE), hence the term *'M'*-estimetion (Fox, 2002).

Consider the linear model

$$y_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \varepsilon_i \qquad i = 1, \ldots, n \tag{8.1}$$

where the $V(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$. If ϵ_i has density f, we can define $\rho = -\log f$, where the function ρ gives the contribution of each residual to the objective function. Then the MLE $\hat{\boldsymbol{\beta}} = \boldsymbol{b}$ solves

$$\min_{\boldsymbol{\beta}} \sum_i -\log f\left(y_i - \mu_i\right) = \min_{\boldsymbol{\beta}} \sum_i \rho\left(y_i - \mu_i\right) \tag{8.2}$$

where $\mu_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}$ and so $\hat{\mu}_i = \boldsymbol{x_i}^\top \boldsymbol{b}$.

Let $\psi = \rho'$ be the derivative of ρ. Then we will have $\sum_i \psi\left(y_i - \hat{\mu}_i\right) \boldsymbol{x_i}^\top = \mathbf{0}$ or $\sum_i w_i \left(y_i - \hat{\mu}_i\right) \boldsymbol{x_i}^\top = \mathbf{0}$ where the weight $w_i = \psi\left(y_i - \hat{\mu}_i\right) / \left(y_i - \hat{\mu}_i\right)$. This suggests an iterative method of solution, updating the weights at each iteration (Venables and Ripley, 2002).

If $\rho\left(x\right) = x^2$, the solution is the conditional mean and the median is $\rho\left(x\right) = |x|$. The function

$$\psi\left(x\right) = \begin{cases} -c & x < -c \\ x & |x| < c \\ c & x > c \end{cases} \tag{8.3}$$

is known as *Winsorizing* and brings in extreme observations to $\mu \pm c$. The corresponding function $\rho = -\log f$ is

$$\rho\left(x\right) = \begin{cases} x^2 & if\, |x| < c \\ c\left(2\left|x\right| - c\right) & \text{otherwise} \end{cases} \tag{8.4}$$

and equivalent to a density with a Gaussian centre and double-exponential tails. This estimator is due to Huber. Note that its limit as $c \to 0$ is the median, and as

$c \to \infty$ the limit is the mean. The value c =1.345 gives 95% efficiency at the normal (Venables and Ripley, 2002).

Venables and Ripley's `MASS` package (1998) introduces the `rlm` function for fitting a linear model by iterated re-weighted least squares (IWLS) regression using Huber's M-estimator with tuning parameter c =1.345 and also incorporating a robust estimate of the scale parameter σ, where $\hat{\sigma} = s$. The details are; if we assume a scaled pdf $f(e/\sigma)/\sigma$ for ϵ and set $\rho = -\log f$, in this case the MLE minimizes

$$\min_{\boldsymbol{\beta}} \left[\sum_i \rho \left(\frac{y_i - \mu_i}{\sigma} \right) + n \log \sigma \right] \tag{8.5}$$

Assuming that σ is known and if $\psi = \rho'$, the MLE b of β solves

$$\min_{\boldsymbol{\beta}} \sum_i \boldsymbol{x}_i \psi \left(\frac{y_i - \mu_i}{\sigma} \right) = 0 \tag{8.6}$$

A common way to solve the above equation is by IWLS, with weights

$$w_i = \psi \left(\frac{y_i - \hat{\mu}_i}{\sigma} \right) \Big/ \left(\frac{y_i - \hat{\mu}_i}{\sigma} \right) \tag{8.7}$$

Of course, in practice the scale σ is not known. However, as mentioned above σ is estimated by a robust MLE-type estimate denoted by s.

A cubic polynomial using the `rlm` function in `R` has been fitted to the *log*-transformed data (in a bid to stabilize the variance over age) using MM-estimation that combines the resistance and robustness, whilst gaining the efficiency of M-estimation.

```
> library(MASS)
> mp<-rlm(log(weight)~1+agey+I(agey^2)+I(agey^3), method="MM")
> summary(mp)

Call: rlm(formula = log(weight) ~ 1 + agey + I(agey^2) +
                          I(agey^3),  method = "MM")
Residuals:
      Min        1Q    Median        3Q       Max
-0.784423 -0.098632 -0.001707  0.096245  0.708137

Coefficients:
            Value    Std. Error t value
(Intercept)   1.1731   0.0034   342.0763
agey          2.0866   0.0148   140.8422
I(agey^2)    -1.1875   0.0145   -81.6688
I(agey^3)     0.2223   0.0036    61.0781

Residual standard error: 0.144 on 6123 degrees of freedom
```

After fitting this cubic line we have used the weights produced in a robust regression procedure to identify the most extreme values. The observations with the big residuals are down weighted, which reflects that they are atypical from the rest of the observations when it comes to fitting such a model. Observations with 0 weight ($w_i = 0$) are deemed to be extreme and so are then removed from the data before running the LMS model fitting algorithm (Figures 8.1 and 8.2). Please note that weight referred to in the Figure 8.1 corresponds to girls actual body weight.

The World Health Organisation (WHO) has defined limits for acceptable data based on 1977NCHS/WHO growth charts and recommends that the exclusion range for weight-for-age should be $|z| > 5$ (17), where z is a z-score of a given measurement. After the final LMS model for girls weight (age birth to three) was fitted, we used the `zscores` function from the `lmsqreg` package in `R` to calculate z-scores for the four identified outliers and these are given in Table 8.1. Each omitted case has an $|z|$

R Data Editor

row.names	weight	agey	w
811	15.7	0.6	0
1235	2.5	0.2617	0
2240	4.8	1.21	0
5963	30.2	2.96	0
1149	11	0.33	0.002548397
1993	6.8	2.7819	0.00273806
3250	20	1.86	0.02485549
4170	8	0.15	0.02486067
1201	9.8	0.3	0.06277497
1305	8.4	0.2	0.06360398
3454	14	0.65	0.1121188
86	18.5	1.97	0.1242005
4131	3	0.259	0.1309751
3595	7	0.1206	0.1326621
5740	6	1.0648	0.1556946
5392	10	0.36	0.1615647

Figure 8.1: *Identifying the outliers, Girls Weight, age birth to 36 months*

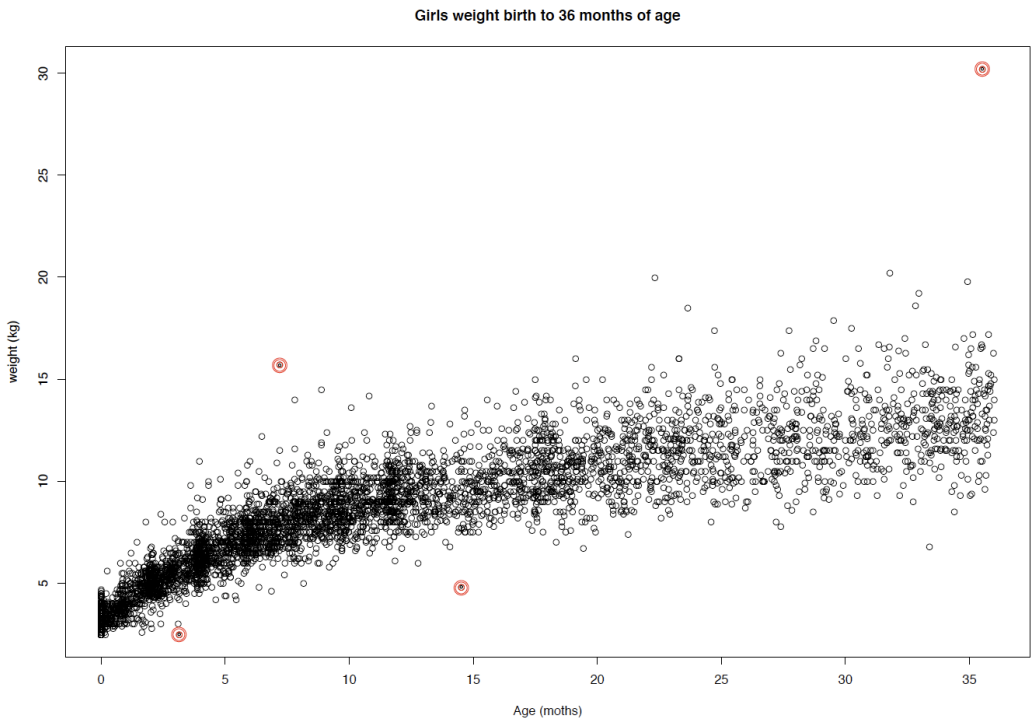


Figure 8.2: *Identified outliers, Girls Weight, age birth to 36 months*

greater than 5 tying in with the WHO guideline.

z-scores {lmsqreg}			
row.names	weight	age	z-score
811	15.70	0.60	5.22488
1235	2.50	0.26	-6.70038
2240	4.80	1.21	-5.29738
5963	30.20	2.96	6.49793

Table 8.1: z-scores of the four identified outliers for girls weight, age birth to 36 months

8.3 LMS

Under the assumption of normality, growth curves can be constructed by estimating the age specific mean and standard deviation, say $\mu(t)$ and $\sigma(t)$, so that chosen quantile curve for $\alpha \in [0, 1]$ can then be obtained as

$$\hat{Q}\left(\alpha \mid t\right) = \hat{\mu}(t) + \hat{\sigma}(t)\Phi^{-1}(\alpha) \tag{8.8}$$

where $\Phi^{-1}(\alpha)$ denotes the inverse of the standard normal distribution function. Providing that assumption of normality holds at each age, such a curve should split the population into two parts with the proportion α lying below the curve, and the proportion of $1 - \alpha$ above the obtained curve (Wei et al., 2006).

Although adult heights in a reasonably homogeneous population are known to be quite close to normal, in general anthropometric data are known to be not normally distributed (Wei et al., 2006). Anthropometry tends to be right skewed rather than left skewed, which is why a log transformation which treats the two tails of the distribution differently is often suggested as a means of obtaining a symmetric distribution (Cole, 1990). A log transformation can be viewed as a particular power transformation of the data but there is a whole family of such powers. Cole (1988) suggested that in principle, there is no reason why a general power transformation should not be applied to the data. The maximum likelihood estimate (MLE) for the power, which both minimises the skewness and optimises the fit to normality, is ideally suited to the problem of skew data. However, it only operates on individual

groups and does not allow for the skewness to change in a smooth manner over the range of the covariate.

The LMS, or $\lambda\mu\sigma$, approach of Cole (1988) provides a way of obtaining normalised growth centiles that deals quite generally with skewness as well as non-constant variance. The method enables us to fit the growth standards to all forms of anthropometry by making the simple assumption that the data can be normalised by using a smoothly varying Box-Cox transformation.

If we denote a variable of interest as Y and assume that it is positive, with mean μ, and that Y^λ is normally distributed, we can consider the transformed variable

$$X = \frac{(Y/\mu)^\lambda - 1}{\lambda}, \qquad \lambda \neq 0 \tag{8.9}$$

or in the case of $\lambda \to 0$ we would have $\ln(Y)$, thus

$$X = \ln\left(\frac{Y}{\mu}\right), \qquad \lambda = 0$$

based on the family of transformations suggested by Box and Cox (1964). This transformation maps the μ of Y to $X = 0$ and is continuous at $\lambda = 0$. For $\lambda = 1$ the standard deviation (σ) of X is exactly the coefficient of variation (CV) of Y and is approximately true for all λ. Minimising the standard deviation of X enables the identification of the optimal λ (Cole and Green, 1992). Transforming the original data by Box-Cox power transformation (8.9) we can define the standardised Z score of X, hence of Y as

$$Z = \frac{X}{\sigma} \tag{8.10}$$

$$= \frac{(Y/\mu)^\lambda - 1}{\lambda\sigma}, \qquad \lambda \neq 0, \tag{8.11}$$

or

$$Z = \frac{\ln(Y/\mu)}{\sigma}, \qquad \lambda \to 0.$$

Enabling the distribution of the measurements Y to varies with covariate t, after the transformation of $Y(t)$ to their standardised values $Z(t)$ they will be normally distributed

$$Z(t) = \frac{[Y(t)/\mu(t)]^{\lambda(t)} - 1}{\lambda(t)\sigma(t)}, \qquad \lambda(t) \neq 0, \tag{8.12}$$

or

$$Z(t) = \frac{\ln[Y(t)/\mu(t)]}{\sigma(t)}, \qquad \lambda(t) \to 0.$$

With these normalised measurements, the desired quantile curve for $\alpha \in [0, 1]$ can then be obtained using the following model

$$Q\left(\alpha \mid t\right) = \mu(t)\left[1 + \lambda(t)\sigma(t)\Phi^{-1}(\alpha)\right]^{1/\lambda(t)} \tag{8.13}$$

which summarises the construction of the centiles by three smooth curves, ie. functions, representing the skewness, the median and the coefficient of variation. As we have seen, the LMS method works with power transformed measurements, but coverts the mean back to original units and uses coefficient of variation (CV) rather than standard deviation of the data. In this way the results for different power transformations can be compared, and the best (Box-Cox) power can be identified as the one which gives the smallest CV (Cole, 1990). This method provides a coherent set of smoothed centiles and the shape of the power curves provide information about the changing skewness, median and coefficient of variation of the distribution.

The three parameters λ, μ and σ were assumed to change smoothly with age. Green (1987) has proposed to estimate the three curves by maximizing the *penalised likelihood*. Let us assume that we have the responses $\{Y_i \colon i = 1, 2, \ldots, n\}$ that are corresponding to design points $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_n$ of a univariate explanatory variable t. We suppose that Y, but not t is subject to error, and we consider the following regression model

$$Y_i = g(t_i) + e_i, \qquad i = 1, 2, \ldots, \ n,$$

where the error is iid with $e_i \sim \mathcal{N}(0, \sigma^2)$, and $g(t)$ is the regression function. Using maximum likelihood we can obtain the estimate for $g(t)$, but most likely the estimate $\hat{g}(t)$ is going to be a rough potrayal for the true dependance of Y on t. In general, maximum likelihood often provides description of the functional dependences that are too complicated, varying too rapidly.

The *penalised likelihood* solution to this difficulty replaces the log-likelihood for the problem

$$\ell(g) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - g(t_i)\right)^2$$

by a new objective function

$$\ell(g) - \frac{1}{2}\nu J(g),$$

where $J(g)$ is a non-negative, real-valued functional measuring the *roghness* of g. Small values of $J(g)$ correspond to function g that are smooth, and vice versa, large values of $J(g)$ correspond to the one that is rough and complex. Thus, when the penalised likelihood is maximised over g, there will be a trade-off between fidelity to the data (large values of $\ell(g)$) and smoothness of the function (small values of $J(g)$) (Green, 1996).

Using penalised likelihood, LMS method enables the three curves for λ, μ and σ to change smoothly with age

$$\ell(\lambda, \mu, \sigma) - \nu_\lambda \int (\lambda''(t))^2 dt - \nu_\mu \int (\mu''(t))^2 dt - \nu_\sigma \int (\sigma''(t))^2 dt \tag{8.14}$$

where $\ell(\lambda, \mu, \sigma)$ is the Box-Cox log-likelihood function derived from (8.12),

$$\ell(\lambda, \mu, \sigma) = \sum_{i=0}^{n} [\lambda(t_i) \log \frac{Y(t_i)}{\mu(t_i)} - \log \sigma(t_i) - \frac{1}{2} Z^2(t_i)] \tag{8.15}$$

and $Z(t_i)$ are the SD scores corresponding to $Y(t_i)$. In this way, the three curves are constrained to change smoothly as the covariate changes and, like the centiles, they can be plotted against the covariate (Figures 8.3 and 8.4). The curves are fitted using cubic splines to give a non-linear regression, and the extent of the smoothing required can be expressed in the terms of smoothing parameters $(\nu_\lambda, \nu_\mu, \nu_\sigma)$. These quantities are defined to be the traces of the relevant smoothing matrices and are referred to as the "equivalent degrees of freedom" (edf) (Wei et al., 2006). Cole and Green (1992) argued that the distributions of $(\nu_\lambda, \nu_\mu, \nu_\sigma)$ in the LMS model are largely independent of each other, implying that one edf can be optimised while fixing the other two.

```
> mw3<-lmsqreg.fit(weight, age, edf=c(7,13,9),
   pvec = c(0.03, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.97))
> plot(mw3)
> points(age, weight, pch=".",col="red")
```

Carey (2002) has developed the `lmsqreg` package that implements the LMS method in `R`. Smoothed centiles curves have been fitted to the reference data using `lmsqreg.fit`

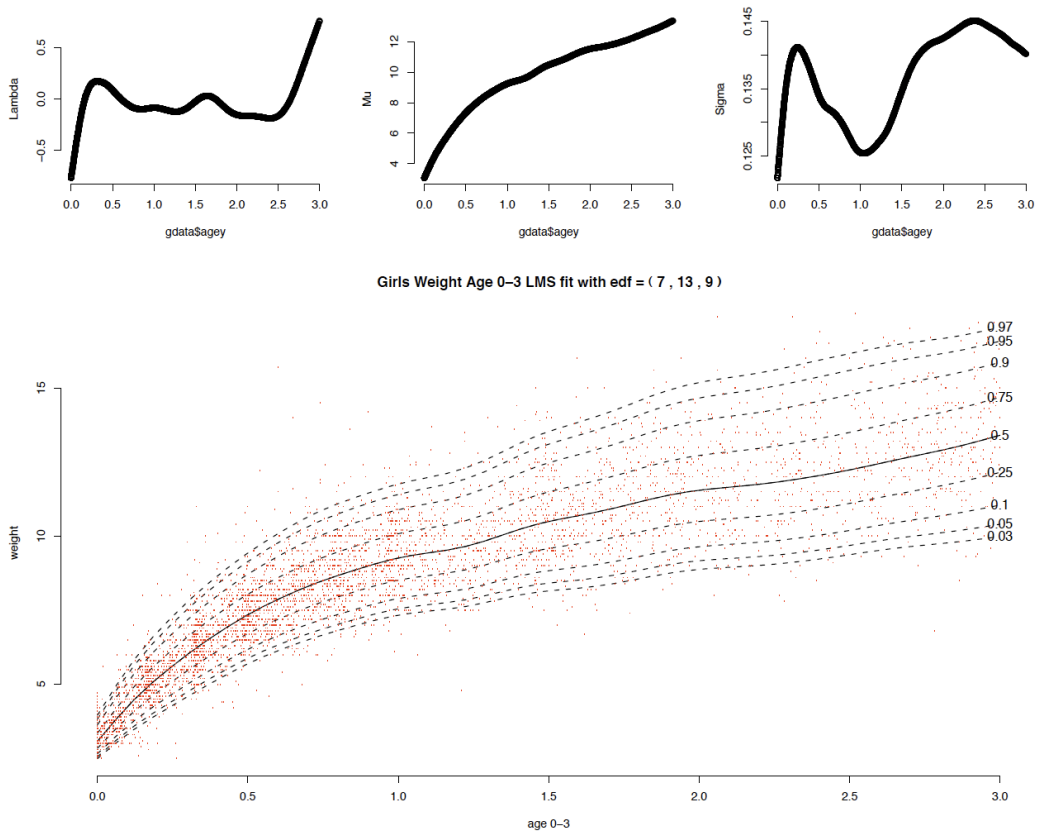


Figure 8.3: *Centile curves for girls weight birth to 36 months of age*

function with suggested starting edf values setting of 3, 5 and 3 for λ, μ and σ, respectively (Carey et al., 2004). The strategy is then to optimise the μ curve edf, by increasing/decreasing the edf by 1 until the change in penalised likelihood is small i.e. less than 2. Once the μ curve is fitted, the process is repeated or the σ curve avoiding the value for edf of 2 which would force a linear trend on the μ curve. Finally, the λ curve was fitted similarly to the σ curve (Figure 8.3). However, in cases of fitting the centiles curves for weight measurement age 2 to 19 years for both sexes λ had to be set to the value of zero, which constrains the entire curve to be a constant value and forces a log transformation (Figure 8.4). The same had to be applied for the fitting of male head circumference age 2 to 19 years.

```
> mw19<-lmsqreg.fit(weight, age, edf=c(0,14,8), pvec = c(0.03,
   0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.97), lam.fixed=0)
> plot(mw19)
```

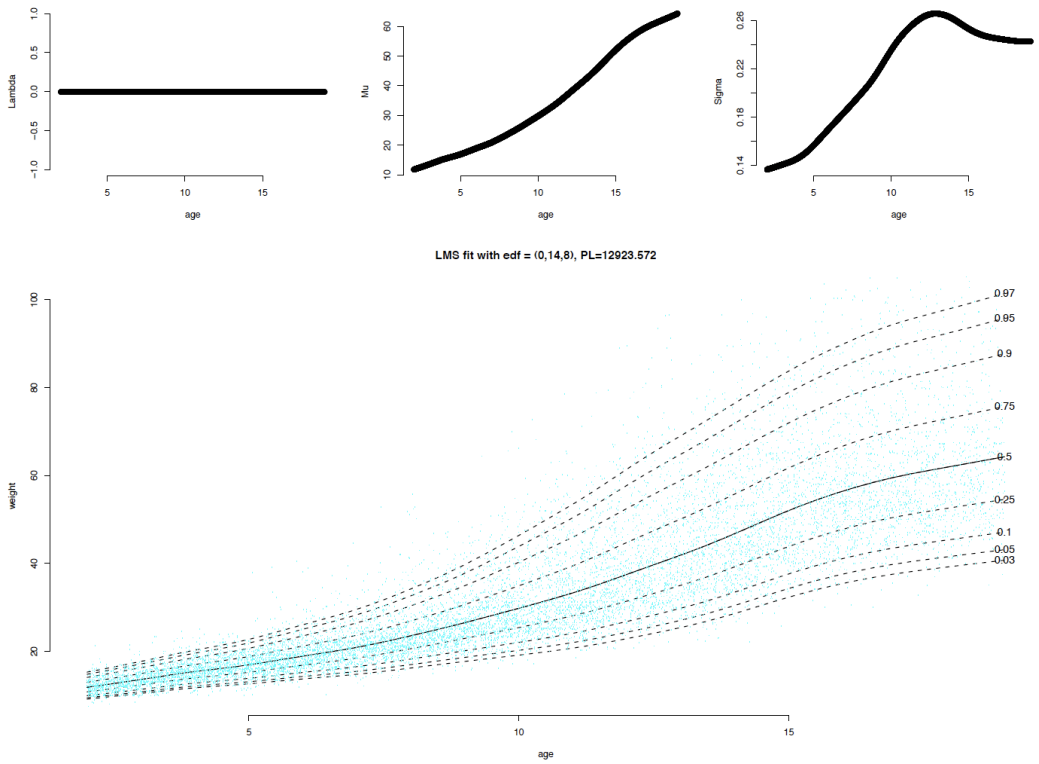


Figure 8.4: *Centile curves for boys weight 2 to 19 years of age*

```
> points(age, weight, pch=".", col="cyan")
```

Following the suggested strategy, the data was over-fitted and the curves were clearly undersmoothed. As Cole (1992) implies the case for making the centile curves smooth is to some extent cosmetic - the centiles are more pleasing to the eye when smoothed appropriately but it is also in the belief that the true population centiles will themselves change smoothly. Any non-parametric curve estimation method requires some means of controling the smoothness of the fitted functions. For the LMS method this control is provided by the edf parameters $(\nu_\lambda, \nu_\mu, \nu_\sigma)$.

As indicated by Carey (2002) the value in which to increase/decrease edf and the change in penalised likelihood depends on the sample size. For large samples the change of less than 2 units is not significant therefore the large change is needed and the final decision should depend on the appearance of the curve. In order to overcome the over-fitting of the curves the edf values had to be relaxed.

8.4 Smoothing and evaluation

The number of effective degrees of freedom is a convenient parameter that expresses the amount of adjustment necessary for smoothing a set of data. Adjustment of edf values was done following Carey's (2002) algorithm, this time decreasing the value for ν_μ by 1 until the curve appeared to be smooth. The same procedure was followed for ν_σ and lastly for ν_λ (Figure 8.5). Finally, the adequacy of the chosen model is evaluated using the original data.

As discussed by Green (1987), the distribution theory for model evaluation statistics formed on the bases of changes in penalised likelihood is currently still undeveloped. We have adopted a local-test based approach to formal model evaluation. Carey's (2002) *lmsqreg* package provides as a part of the output for a fitted model a collection of model-based z-scores derived from the given quantile regression model. They are stratified based on the covariate t, and within this strata, z-scores are tested for marginal Gaussianity (Kolmogorov-Smirnov test), zero mean (Student's t-test) and unit variance (χ^2 test) (15).

```
> mw3
Dependent variable: gdata$weight , independent variable: gdata$agey
The fit converged with EDF=( 4,6,3 ), PL= 9198.316

KS tests: (intervals in gdata$agey //p-values)
   (-0.001,0]     (0,0.348] (0.348,0.802]  (0.802,1.54]      (1.54,3]    Overall
        0.000         0.000         0.271         0.324         0.676      0.001

t tests: (intervals in gdata$agey //p-values)
   (-0.001,0]     (0,0.348] (0.348,0.802]  (0.802,1.54]      (1.54,3]    Overall
        0.006         0.000         0.562         0.369         0.568      0.810

X2 tests (unit variance): (intervals in gdata$agey //p-values)
   (-0.001,0]     (0,0.348] (0.348,0.802]  (0.802,1.54]      (1.54,3]    Overall
        0.000         0.000         0.717         0.050         0.462      0.979
```

The above output from the final fitted model shows that the hypotheses of a zero mean, unit variance normal distribution in the intervals close to birth are rejected. The original data is strongly skewed and the edf parameters finally selected are not able to transform the data sufficiently well, with the final empirical distribution being

slightly skewed. If the smoothing parameters are increased, in particular ν_λ, the normality of the transformed data can be successfully achieved. However, as discussed earlier in section 3, we reduced the values of the optimal smoothing parameters in order to obtain smoother estimated centiles curves.

Table 8.2 reports on the accuracy of the quantile regression fit in terms of the discrepency between the nominal and empirical proportions of data lying beneath selected quantile function for age group birth to 3 years. By and large these results show that the quantiles of the fitted models do fit the data well.

			p								
sex	variable	N	0.03	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.97
female	weight	6,123	0.025	0.052	0.090	0.240	0.506	0.755	0.905	0.950	0.972

Table 8.2: Table entries are quantile coverage probability estimates. Measurement: Age: birth to 36 months.

8.5 Averaging

We were required to produce reference standards for two age groups: birth to 36 months of age and 2 to 19 years of age. The overlap for the two sets of charts occurs for ages between 2 and 3 years. The values for both sets of standards in the overlapping age range is a product of the model fitted to the whole data set for each specific age group. This means that the centile curves for a particular measurement in this overlapping period will not be the same for the two sets of charts as they are based on using different data outside the range 2 to 3 years (Figure 8.6).

One of the arguments of `lmsqreg.fit` function is `targlen` which defines the number of points at which smooth estimates of λ, μ, and σ should be extracted for quantile plotting. For both sets of charts we have adopted the default value of 50 for the `targlen` argument. For the overlapping period 2 to 3 years this produces 17 points in the birth to 36 months chart and 3 points in the chart for age 2 to 19 years (Figure 8.7).

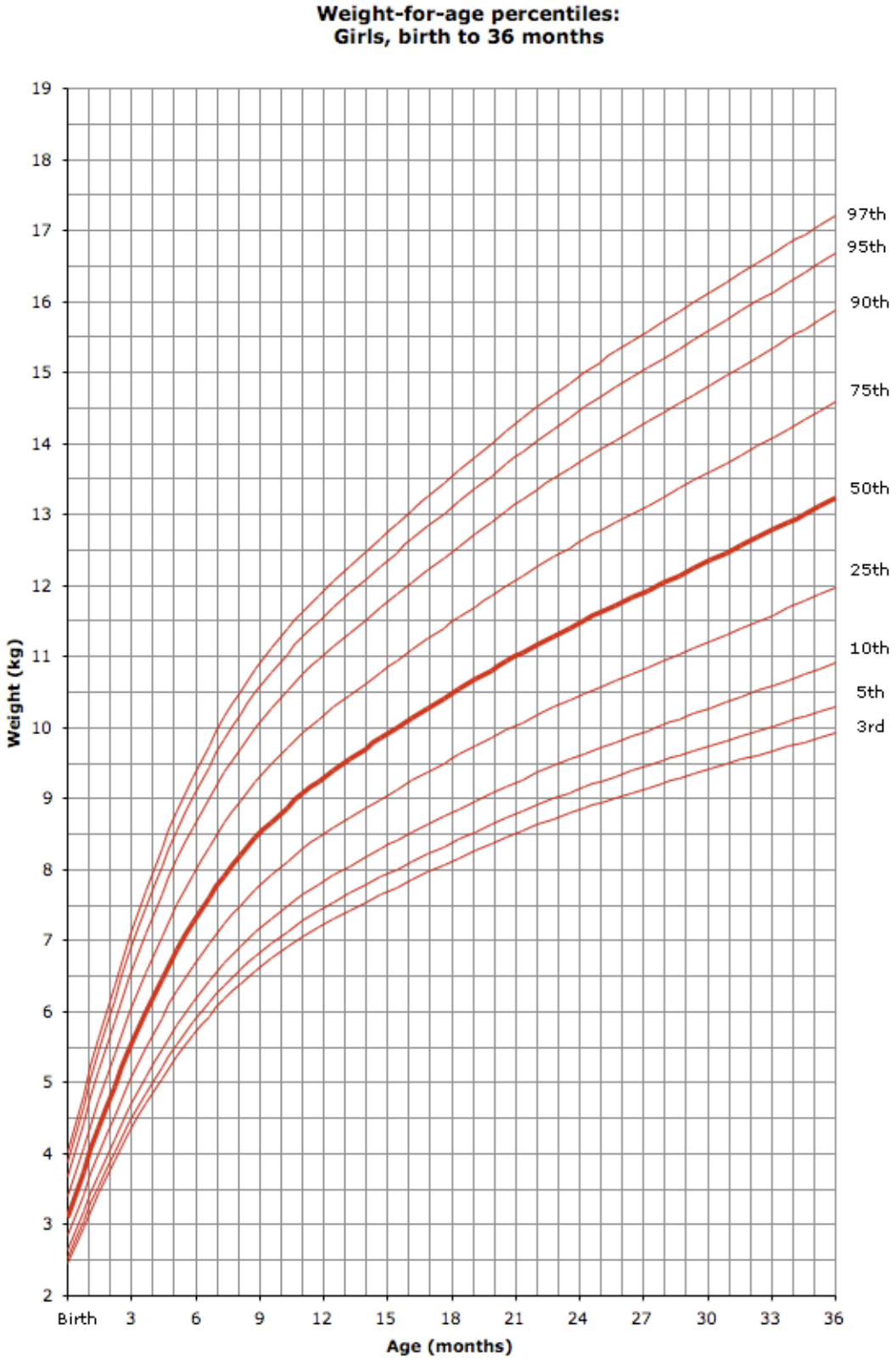


Figure 8.5: *Final smooth centile curves for girls weight birth to 36 months of age*

Figure 8.6: *Overlapping charts: girls weight*

In order to make the centile curves for a particular measurement for this overlapping period the same for the two sets of charts we have re-estimated the curves using the following cubic polynomial:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 \tag{8.16}$$

To estimate this cubic polynomial for each of the centiles at the lower and upper boundaries of the overlapping period we have used three adjacent points from each of the charts (Figures 8.7 and 8.8), using the least squares estimator given by (8.17).

$$\hat{Y} = X \left[X^\top X\right]^{-1} X^\top Y \tag{8.17}$$

For the overlapping period new estimates were calculated using the newly found polynomial resulting in a smooth overlap (Figure 8.9). This means that the centiles for a particular measurement will be the same in the birth to 36 months chart as in the 2 to 19 years age chart.

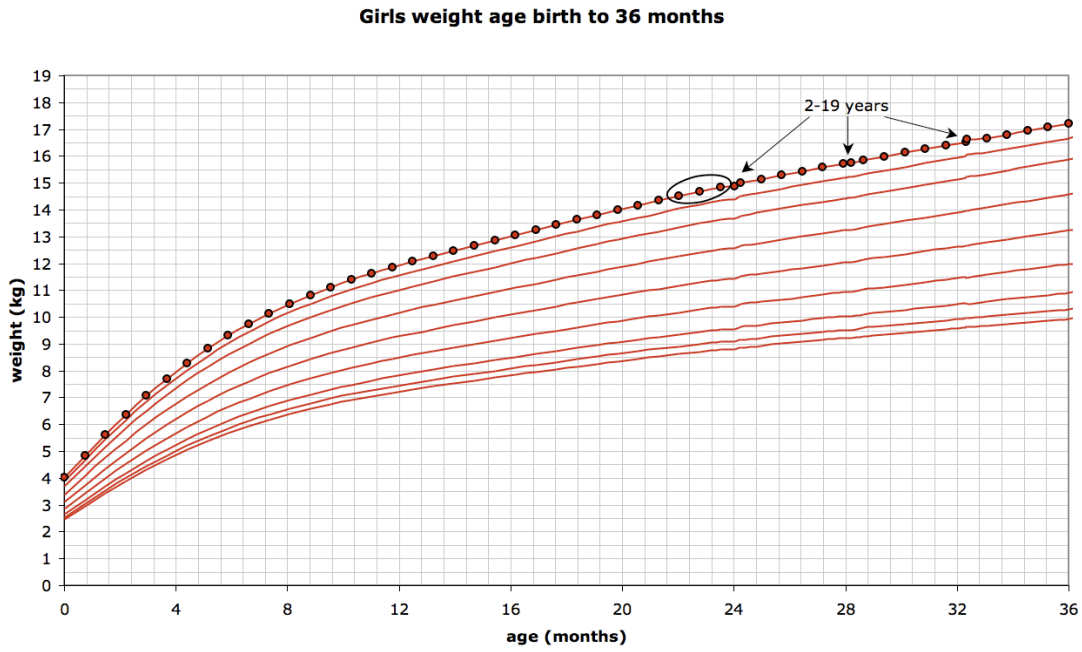


Figure 8.7: *Centile curves for girls weight birth to 36 months of age*

8.6 Comparisons using ANCOVA

8.6.1 Comparing Geographical Regions

In the following analysis the aim is, for a particular measurement, sex and age group, to compare the growth trends over age in different geographical regions. These are:

i. North,

ii. Southwest,

iii. Central.

This means that we are looking at a large proportion of the original data used to fit the LMS models but not all of it as some individuals live in regions other than those listed above.

One approach, for a particular measurement and sex, would be to fit a different LMS model to the data in each region and then to compare the fitted models. However, we are not aware of any existing methodology to make such direct LMS model comparisons. In our proposed approach, we have taken the final LMS model fitted to

Figure 8.8: *Centile curves for girls weight age 2 to 19 years*

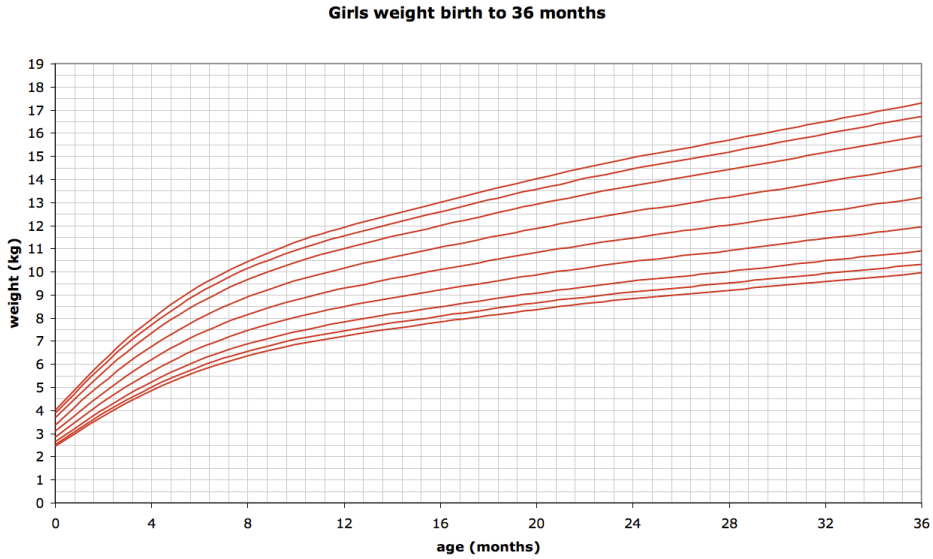


Figure 8.9: *Final smooth centile curves for girls weight birth to 36 months of age*

all the data and used it to transform all the individual measurements into standard deviation scores.

Then, in step 1, a separate cubic regression curve was fitted, where the response ("y-variable") is the SDS score and the covariate ("x-variable") is age, to the data in each of the three regional groups. These regression lines describe how the mean SDS score of a given measurement changes with age in each region. The fit of the three

cubic regression curves were then compared with the fit of three quadratic regression curves. If the difference in fits was not statistically significant then the quadratic models were accepted and they were then compared with three linear regression curves and so on until the simplest model that might be fitted is three different constant horizontal lines. The three final regression lines can be plotted to provide a graphical description of the differences (Figure 8.10).

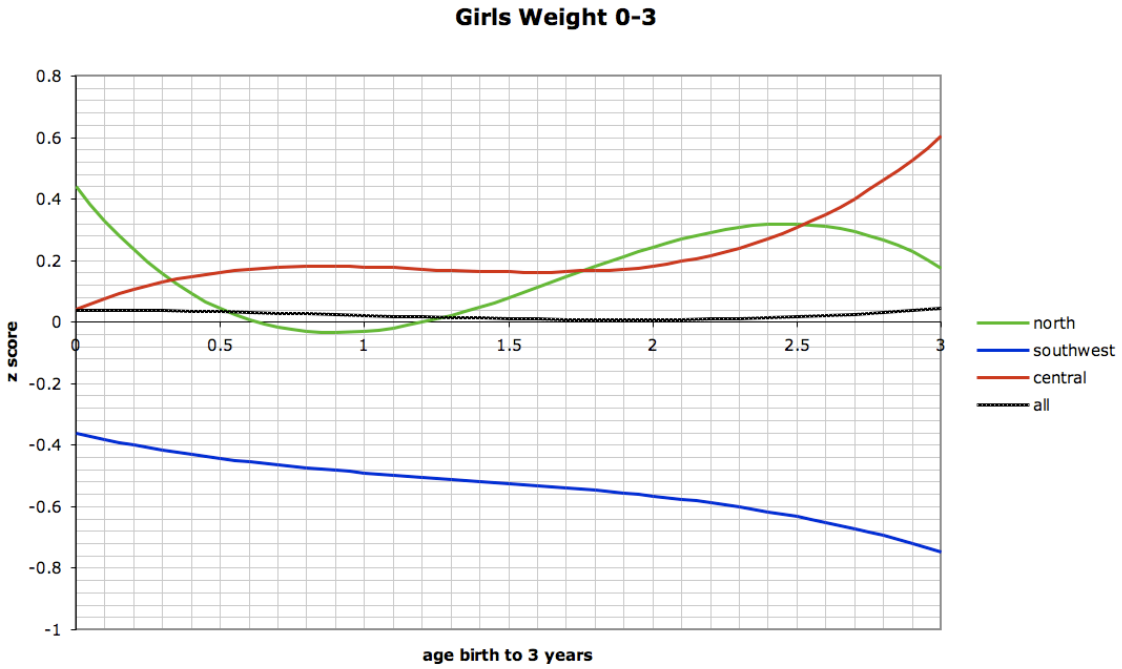


Figure 8.10: *SDS score regression models in the three geographical regions for weight vs. age; age: birth to 36 months; sex: female*

If there are no differences in the three regions in how a particular measurement for a given age group and sex changes with age then a single common regression line would be an appropriate model for all the data in the three regions. Therefore, in step 2, such a model was fitted to the data. It would be expected that it would be fairly close to the zero line but not identically zero because we have not used all the original data in this analysis as explained above. The degree of this line (cubic, quadratic, etc.) was chosen to be the same as that of the best fitting three separate ones.

The next stage is to statistically test the fit of the model involving three separate regression lines with the fit of the model based on a single common regression line. We would expect the total residual sum of the squares for the model involving three regression lines to be less than that which just involves one but we need to test whether the difference is statistically significant. The method we have used is a standard *"F-test"* in this context, which is appropriate because the standardised data is Normally distributed . If the p-value of this test is small (less than 0.05) the conclusion would be that the single regression line is inadequate and there are significant differences between the regions in how mean SDS score of a given measurement of a given age group of a given sex changes with age (Table 8.3).

age: birth to 36 months		
sex	variable	p
female	weight	$< 10^{-6}$

Table 8.3: Resulting p-values when testing a common regression model vs. different regression models for the three regions.

Finally, after finding a significant result we can then go on to use the same methodology as above but just use pairs of regions in turn to see which are significantly different from each other.

This procedure can be summarized for a given sex and measurement by the following steps:

[i] STEP 1: Find the best fitting polynomials having the lowest possible common degree for each of the three regions.

[ii] STEP 2: We want to answer the question "Is a common polynomial of the same degree as found in STEP 1 appropriate for all three regions or do the polynomials vary with region?"
ie. for a particular measurements, sex and age group we want to test:
$H_0 : E[z \mid age] = \beta_0 + ... + \beta_q age^q$ for each region, where $q \leq 3$ is the degree of the common best fitting polynomial.
vs. H_1 : The polynomial for at least two regions differ.

[iii] STEP 3: After finding a significant result in STEP 2 cary out pairwise comparisons between the regions.

Note that the notation $E[z|age]$ denotes the mean value of z at the given age. The hypothesis H_0 says that a common polynomial of degree q describes the trend in z-scores over age in each region. On the other hand, the alternative hypothesis, H_1, says that the trend in z-scores is described by different polynomials of the same degree q in the regions.

Nb. in the following tables p denotes the "p-value" found when testing as above H_0 vs. H_1. Its value corresponds to the probability of observing a test statistic value at least as large as we have done and is calculated under the assumption that the null hypothesis, H_0 , is true. For the stepwise and overall tests in STEP 1 and STEP 2 above it is common practice to reject H_0 in favor of H_1 if $p < 0.05$ or if the sample size is large than we may use $p < 0.01$. To account for carrying out multiple comparisons (or multiple hypothesis tests) between pairs of regions for a particular measure as in STEP 3, we would suggest that H_0 is rejected in favor of H_1 if $p < 0.003$ (ie. 0.01/3) using the Bonferroni method which divides the total significance level into 3 equal proportions corresponding to the number of pairwise comparisons we are carrying out.

Note that the results of the analyses carried out in STEP 2 for age birth to 3 years are given in Table 8.4. The coefficients of the polynomials in the three separate regions, as well as for all three regions together, are in Tables 8.4. Those polynomials for girls weight age birth to 3 years are plotted in Figure 8.10. Table 8.5 details the p-values for all the pairwise comparisons between regions.

sex: female									
variable	region	$\hat{\beta}_0$	se	$\hat{\beta}_1$	se	$\hat{\beta}_2$	se	$\hat{\beta}_3$	se
weight	central	0.042	0.039	0.388	0.169	-0.342	0.173	0.092	0.045
	north	0.441	0.055	-1.208	0.235	0.917	0.230	-0.181	0.057
	southwest	-0.363	0.058	-0.206	0.282	0.106	0.285	-0.026	0.072
	all	0.040	0.029	-0.005	0.127	-0.020	0.128	0.008	0.032

Table 8.4: Estimates of the model parameters for individual regions and all three regions together for female weight, age range birth to 36 months.

There are clearly significant differences between the regions for each of the measurements for both sexes in each of the age ranges.

sex:female, age birth to 36 moths	
weight	
	p
north-central	$< 10^{-6}$
southwest-central	$< 10^{-6}$
southwest-north	$< 10^{-6}$

Table 8.5: p values for the pairwise comparisons between the different regions using ANCOVA

8.6.2 Comparing Males and Females

Standard deviation scores were used to compare the growth patterns between boys and girls using very similar methodology as to that described above when comparing the geographical regions. In order to make comparisons for a given measure between genders, we have used the relevant fitted girls' LMS model to standardise both girls and boys measures using the `zscores` function from Carey's (2002) `lmsqreg` package.

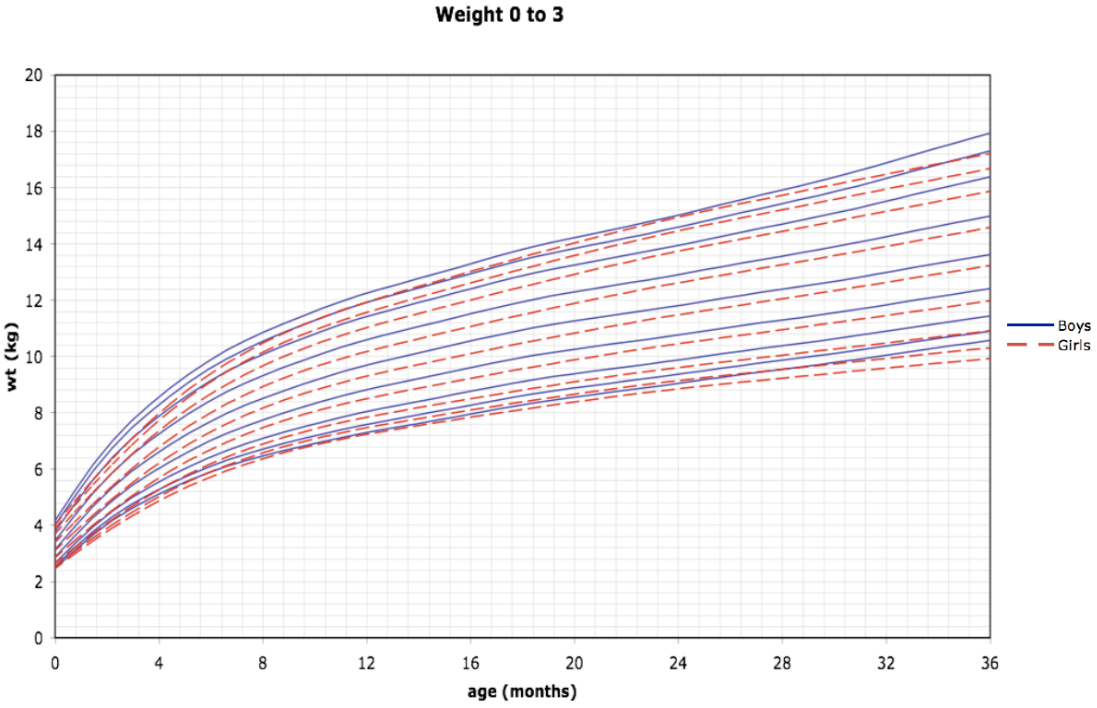


Figure 8.11: *Comparisons of the growth charts for weight measurement between male and female birth to 3 years of age.*

We can then plot these standardised measures against age and construct separate regression lines for boys and girls. Considering that the data were standardised by the

girls model, it is evident that the appropriate regression model for girls would be zero. However, the z scores of the boys could be explained by an appropriate polynomial regression model (up to cubic polynomial), describing the existing differences between boys and girls. If there are differences then this will be indicated by a non-zero regression line and we can test whether the two lines are significantly differnt from each other using ANCOVA. We have also superimposed girls and boys centiles for a given measure on the same plot to give another graphical impression of any differences (Figure 8.11). For children aged 0-3 we found significant differences for each measure and the fitted regression lines (Table 8.6) describe how the differences (measured in girls standard deviation scores) change with age (Figure 8.12).

Age birth to 36 months								
variable	$\hat{\beta}_0$	se	$\hat{\beta}_1$	se	$\hat{\beta}_2$	se	$\hat{\beta}_3$	se
length	0.217	0.021	0.128	0.048	-0.067	0.019	-	-
head circumference	0.223	0.024	0.758	0.104	-0.564	0.103	0.127	0.026
weight	0.168	0.023	0.701	0.100	-0.655	0.099	0.148	0.025
body mass index	-	-	0.525	0.075	-0.505	0.086	0.122	0.022

Table 8.6: Estimates of the model's parameters.

8.7 Conclusion

This study was set up by the Saudi medical authorities who required growth charts based entirely on data collected from Saudi children and adolescents rather than using a more gebneral alternative, such as those provided by the WHO. We have seen that for girls weight (birth to 36 months) the age-specific conditional quantile estimates we have constructed using the LMS method by and large successfully capture the main features of the data and this also proved to be true for the other growth parameters. In further work we have compared the new Saudi charts with the 2006 WHO standards and found that there are marked differences in corresponding centiles. Use of the WHO standards in Saudi Arabia would, for example, increase the prevalence of undernutrition, stunting and wasting (El Muzan et al., 2009).

An essential part of our procedure was to try to identify outliers to be removed

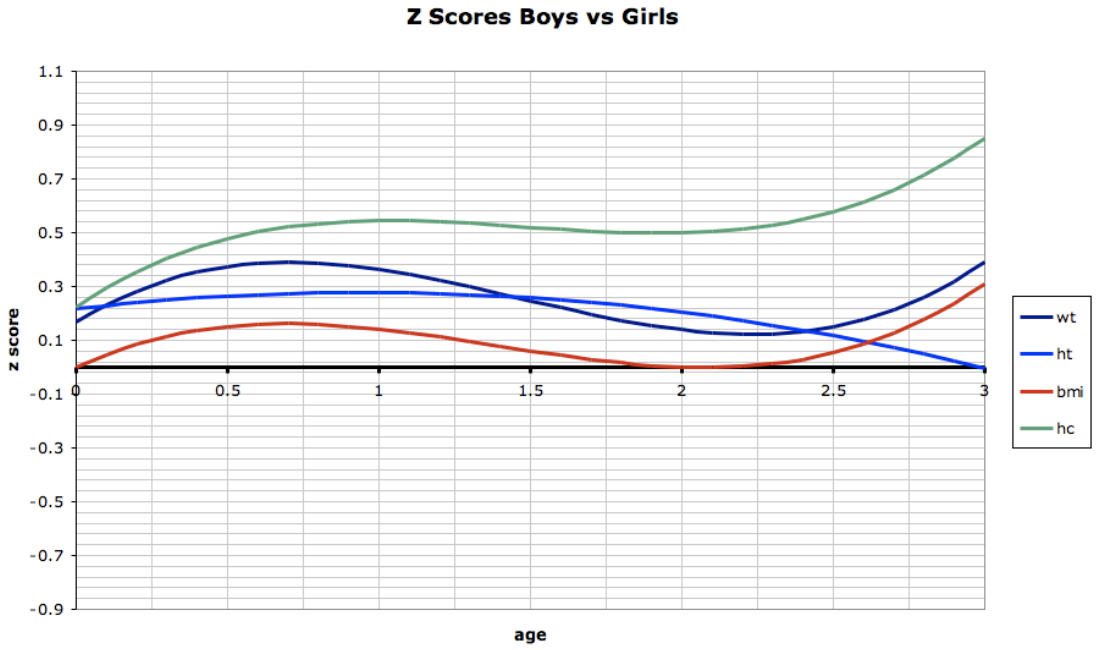


Figure 8.12: *Comparisons of growth patterns between boys and girls birth to 3 years of age*

from the data prior to estimating the LMS model. We used the robust regression rlm function to do this, basing our assessment on the weight attached to each observation by the procedure we should stress that we were not using this model to make any formal inferences about the form of the conditional mean function. As seen in Section 2, this worked well with four cases being removed. If these cases were included then there is a numerical failure in the LMS model estimation algorithm. All four deleted cases had z-scores greater than 5 in absolute value. The only other case which had an absolute z-score bigger than 5 is case 4131 with a z-score of -5.078, who can be seen listed in Figure 8.1. This corresponds to a girl aged 0.259 years (3.11 months) who had a weight of only 3.0 kg which is a little higher than case 1235 whose weight was only 2.5 kg at a similar age and who was deleted from the data.

A number of authors have reported that there can be a problem with significant kurtosis in the residuals from the LMS method (Stasinopoulos and Rigby, 2004). Stasinopoulos and Rigby have developed the more flexible Box-Cox power exponential model to overcome this where thay add an extra parameter to model kurtosis

(Stasinopoulos, Rigby and Akantziliotou, 2009). The LMS model we have used is a special case of this. In this study we did not consider this alternative.

Chapter 9

Conclusions

The kernel smoothing bootstrapping for parametric quantile regression is a new method for constructing the covariance matrix of the parametric estimates. Throughout the study we have been assessing its performance in different data scenarios with the aim to identify an appropriate variance function estimation procedure required by this approach and to examine its robustness.

By modeling the conditional mean of the squared residuals using gamma GLM, we estimated the conditional variance function of the τ^{th} quantile under the assumption of error being normally distributed. In Chapter 4 we explored the robustness of this method when the error was not normal and therefore the residuals are non gamma. We found that *glm.Gamma* was less sensitive to the distribution of the response than expected, however for distributions with long tails, the gamma GLM fitting procedure could fail. In order to overcome this difficulty we proposed to use the *Double Generalised Linear Models*, which model the mean and dispersion simultaneously in the context of a general linear model, allowing for heteroscedastic dispersion (Smyth and Verbyla, 1999a), assuming that the gamma distribution of the centered squared residuals could be mimiced by a family of Tweedie distribution with $p = 2$. Chapters 6 and 7 considered non-parametric approaches used for conditional variance modeling, but recommend that using the DGLM with the Tweedie family with the *log* link is still the preferable option.

Finally, the methodology suggested above is applied to the Saudi Arabian anthropometric measurements using the girls' weight data from birth to 3 years of age. We fited a cubic parametric quantile regression model to the data (see Figure 9.1)

$$\mathcal{Q}_y(\tau|\boldsymbol{x}) = \beta_0(\tau) + \beta_1(\tau)x + \beta_2(\tau)x^2 + \beta_3(\tau)x^3, \quad (9.1)$$
$$\text{for } \tau \in (0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)$$

that corresponds most appropriately to the derived LMS standards shown in Figure 8.5. For the purpose of easy comparisons with the other existing methods we focus only on the 50^{th} and 75^{th} quantiles. The conditional variance function was estimated using the proposed DGLM with the Tweedie family with the *log* link by fitting a cubic model to the centered squared residuals as suggested by the estimated curve of the σ parameter obtained from the LMS method (see Figure 8.3). The results of this are summarised below:

Fitted models using `rq.fit` function in `R`:

$$\begin{aligned}\mathcal{Q}_y(\tau = 0.5|\boldsymbol{x}) &= 3.100 + 10.630x - 5.166x^2 + 0.939x^3,\\ \mathcal{Q}_y(\tau = 0.75|\boldsymbol{x}) &= 3.400 + 11.371x - 5.412x^2 + 0.973x^3.\end{aligned}$$

Parameter estimates by all methods for both quantile functions are given in Table 9.1. The computed 95% confidence intervals (CI) for each of the two models parameters and their lengths (L) are given in Tables 9.2 and 9.3 respectively.

The confidence intervals of the parameter estimates of b_1, b_2 and b_3 in both models are the narrowest when applying the kernel smoothing bootstrapping and the parameters estimates are very close to the estimates from the original models.

The confidence interval of the parameter estimate of b_0 in the both models is the smallest compared to the other bootstrapping methods considered. However, the Wald methods, which accommodate for heteroscedasticity, and the rank-score methods give narrower lengths than kernel smoothing bootstrapping. The estimate of the intercept parameter obtained by the kernel smoothing bootstrapping is slightly inflated in comparison to the others, and the confidence intervals does not include the original parameter estimate. If we look at the Figure 9.2, showing 95% pointwise

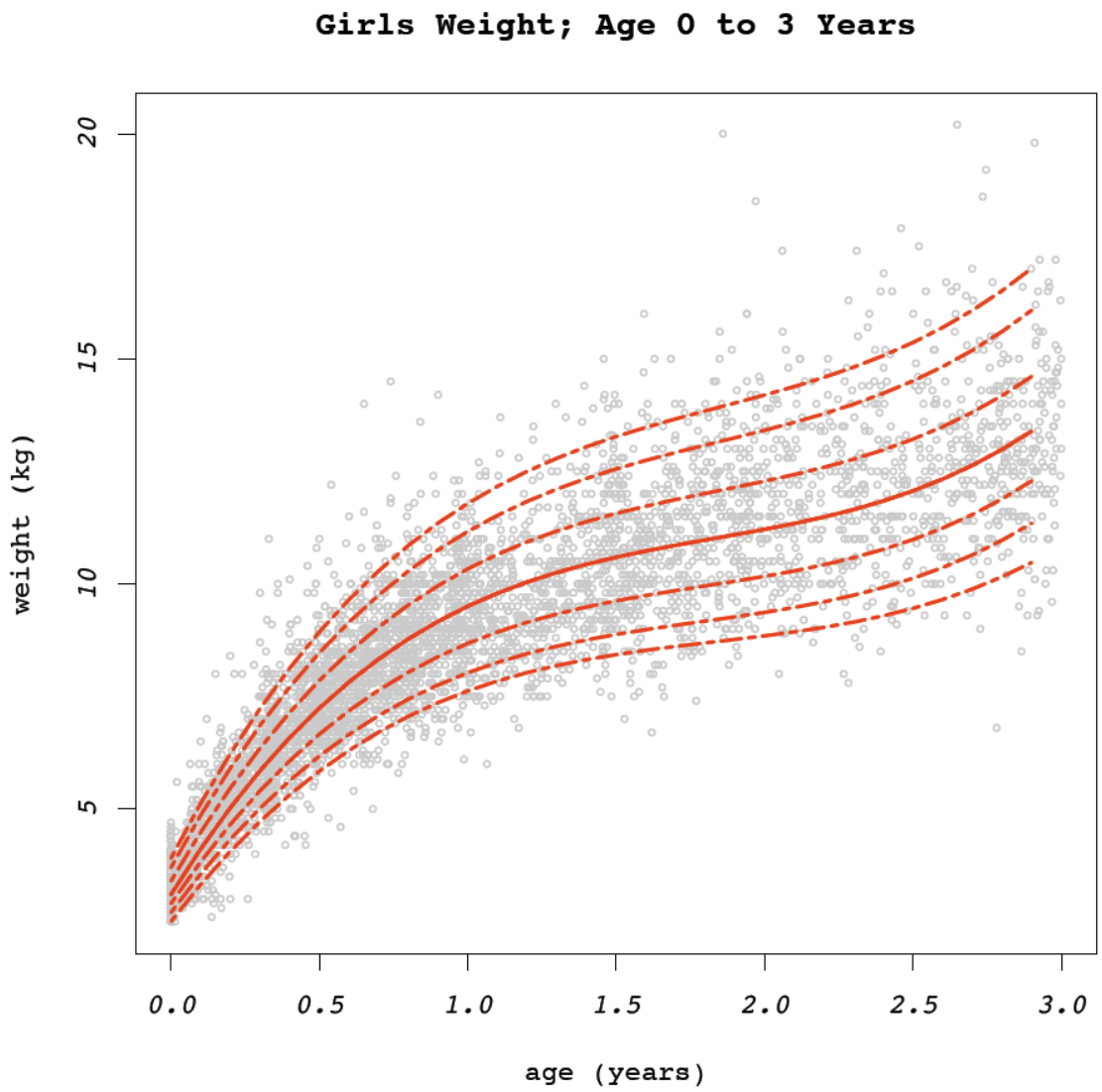


Figure 9.1: *Girls Weight: Scatter diagram of girls weight, age birth to 3 years. Superimposed are fitted cubic quantile regression curves to the data corresponding to* $\tau \in (0,05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)$.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$\tau = 0.50$				
ksm	3.15664	10.62845	-5.16428	0.93900
xy	3.11342	10.56746	-5.11237	0.92739
pwy	3.11213	10.57861	-5.12668	0.93147
mcmb	3.13091	10.50225	-5.06408	0.91734
wxy	3.11211	10.57285	-5.11851	0.92895
riid	3.10000	10.63040	-5.16633	0.93950
rnid	3.10000	10.63040	-5.16633	0.93950
wiid	3.10000	10.63040	-5.16633	0.93950
wker	3.10000	10.63040	-5.16633	0.93950
wnid	3.10000	10.63040	-5.16633	0.93950
$\tau = 0.75$				
ksm	3.47945	11.36891	-5.41055	0.97327
xy	3.41857	11.30777	-5.36420	0.96378
pwy	3.41879	11.31581	-5.37643	0.96711
mcmb	3.43122	11.26144	-5.32934	0.95660
wxy	3.41785	11.31142	-5.37044	0.96564
riid	3.40000	11.37132	-5.41289	0.97385
rnid	3.40000	11.37132	-5.41289	0.97385
wiid	3.40000	11.37132	-5.41289	0.97385
wker	3.40000	11.37132	-5.41289	0.97385
wnid	3.40000	11.37132	-5.41289	0.97385

Table 9.1: Parameter estimates for Model given by (9.1), for $\tau = 0.5$ and $\tau = 0.5$ quantile functions. Methods used for the estimation are: kernel smooth bootstrap (ksm), (x, y) pair bootstrap (xy); Parzen-Wai-Ying bootstrap (pwy); markov chain marginal bootstrap ($mcmb$); generalised bootstrap method of Bose and Chatterjee (2003) with unit exponential weight (wxy); rank-score test method assuming iid errors ($riid$); rank-score test method assuming nid errors ($rnid$); Wald method assuming iid error using Hall and Sheather's bandwidth ($wiid$); Wald method assuming nid error with Powell's sandwich estimate ($wker$); and Wald method assuming nid error, with Siddiqui sandwicj estimate, using Hall and Sheather's bandwidth ($wnid$).

confidence intervals of the estimated quantile functions for $\tau = 0.5$ and $\tau = 0.75$ we can notice that they not fully cover the entire function estimate. Taking a closer look at the the fitted quantile functions shown in Figure 9.1 we notice that there is a number of observations at age zero with larger weights. An initial suggestion is that in estimating the density function at age zero the kernel smoothing bootstrapping is introducing bias towards those observations effectively pulling the location of the intercept towards them. This potential problem should be further explored compared

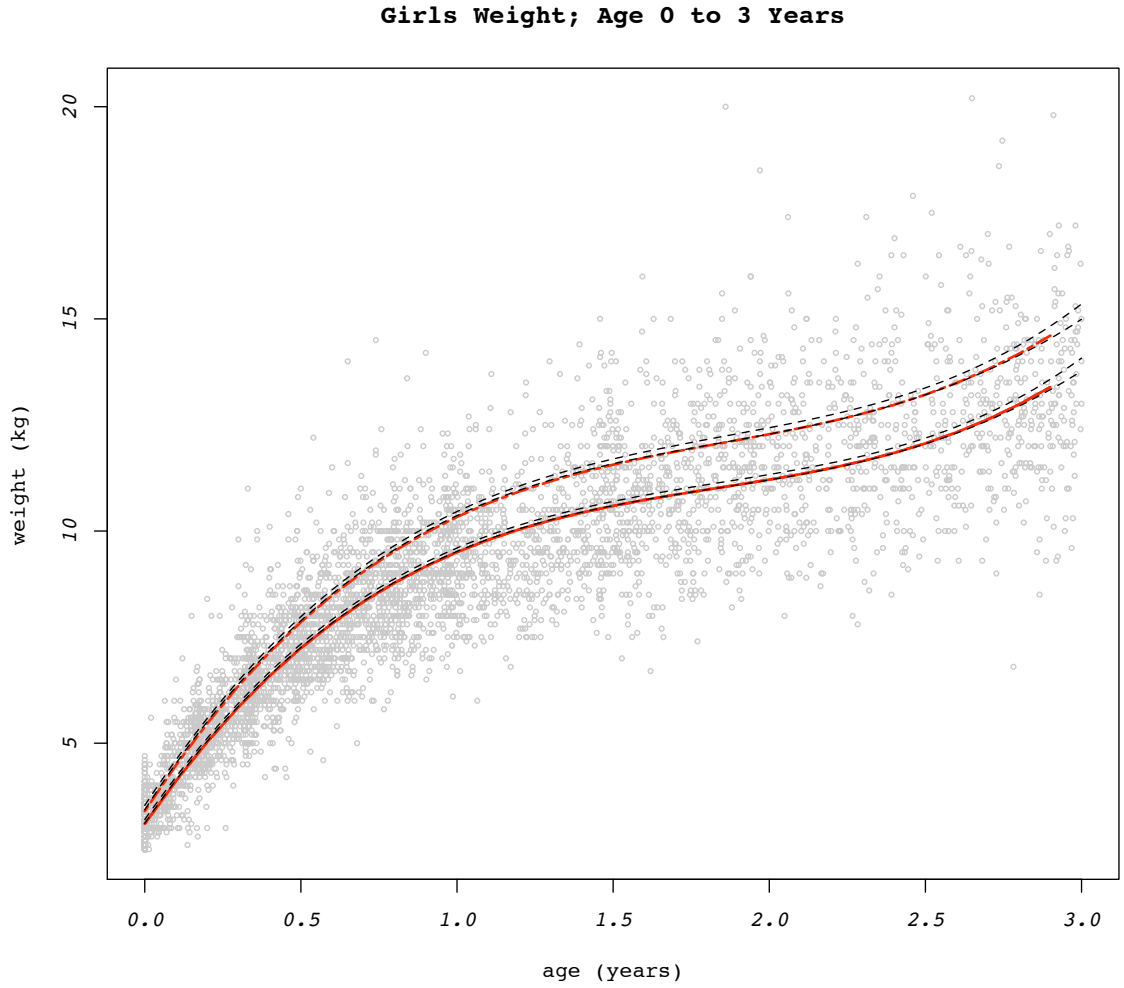


Figure 9.2: *Girls Weight: Scatter diagram of girls weight, age birth to 3 years. Superimposed, solid red lines, are fitted cubic quantile regression curves to the data corresponding to* $\tau \in (0.50, 0.75)$. *The 95% pointwise confidence intervals of the two estimated quantile regression curves are presented by the dashed lines.*

	β_0	β_1	β_2	β_3
Method	CI	CI	CI	CI
$\tau = 0.50$				
ksm	(3.10939, 3.20390)	(10.42505, 10.83186)	(-5.36327, -4.96530)	(0.88927, 0.98873)
xy	(3.05842, 3.16840)	(10.28178, 10.85314)	(-5.42069, -4.80405)	(0.84266, 1.01211)
pwy	(3.06065, 3.16361)	(10.29828, 10.85894)	(-5.42727, -4.82609)	(0.84959, 1.01335)
mcmb	(3.06887, 3.19295)	(10.20688, 10.79761)	(-5.37718, -4.75099)	(0.83341, 1.00127)
wxy	(3.05992, 3.16430)	(10.27896, 10.86674)	(-5.44358, -4.79344)	(0.84036, 1.01755)
riid	(3.10000, 3.18391)	(10.24849, 10.72993)	(-5.32348, -4.84358)	(0.87184, 0.98763)
rnid	(3.10000, 3.18495)	(10.22061, 10.74803)	(-5.33368, -4.83012)	(0.86487, 0.98902)
wiid	(3.02353, 3.17648)	(10.29992, 10.96088)	(-5.49078, -4.84188)	(0.85825, 1.02074)
wker	(3.07249, 3.12751)	(10.39854, 10.86226)	(-5.44893, -4.88374)	(0.85983, 1.01915)
wnid	(3.07624, 3.12376)	(10.40658, 10.85422)	(-5.44256, -4.89011)	(0.86118, 1.01781)
$\tau = 0.75$				
ksm	(3.42662, 3.53226)	(11.14117, 11.59665)	(-5.63411, -5.18700)	(0.91735, 1.02920)
xy	(3.36006, 3.47708)	(11.00136, 11.61418)	(-5.68867, -5.03973)	(0.87810, 1.04945)
pwy	(3.35644, 3.48113)	(10.99943, 11.63219)	(-5.70829, -5.04457)	(0.88022, 1.05400)
mcmb	(3.36785, 3.49459)	(10.94250, 11.58038)	(-5.66794, -4.99073)	(0.86727, 1.04592)
wxy	(3.36006, 3.47564)	(11.01214, 11.61069)	(-5.68443, -5.05646)	(0.88346, 1.04782)
riid	(3.40000, 3.50101)	(11.00741, 11.51621)	(-5.61632, -5.12840)	(0.89823, 1.03332)
rnid	(3.40000, 3.50245)	(11.00325, 11.51826)	(-5.63085, -5.11649)	(0.89375, 1.03355)
wiid	(3.31883, 3.48117)	(11.02053, 11.72212)	(-5.75728, -5.06850)	(0.88762, 1.06009)
wker	(3.36719, 3.43281)	(11.12392, 11.61873)	(-5.70444, -5.12134)	(0.89295, 1.05475)
wnid	(3.35814, 3.44186)	(11.10148, 11.64117)	(-5.72937, -5.09641)	(0.88773, 1.05998)

Table 9.2: Computed 95% confidence intervals of the parameters in the models given by (9.1), for $\tau = 0.50$ and $\tau = 0.75$ quantile functions. Methods used for the estimation are: kernel smooth bootstrap (ksm), (x, y) pair bootstrap (xy); Parzen-Wai-Ying bootstrap (pwy); markov chain marginal bootstrap ($mcmb$); generalised bootstrap method of Bose and Chatterjee (2003) with unit exponential weight (wxy); rank-score test method assuming iid errors ($riid$); rank-score test method assuming nid errors ($rnid$); Wald method assuming iid error using Hall and Sheather's bandwidth ($wiid$); Wald method assuming nid error with Powell's sandwich estimate ($wker$); and Wald method assuming nid error, with Siddiqui sandwicj estimate, using Hall and Sheather's bandwidth ($wnid$).

to other bootstrapping methods to try to find out exactly what the nature of this problem is. Further research could be undertaken to investigate the possibility of a bias correction of the density estimation used in the kernel smoothing bootstrapping procedure or whether a more structural response is required.

The kernel smoothing bootstrapping method may be applicable to a range of other statistical models, in particular other types of regression models with heteroscedastic

	β_0	β_1	β_2	β_3
Method	L	L	L	L
$\tau = 0.50$				
ksm	0.09451	0.40680	0.39796	0.09946
xy	0.10998	0.57136	0.61664	0.16945
pwy	0.10296	0.56066	0.60119	0.16377
mcmb	0.12408	0.59073	0.62619	0.16785
wxy	0.10438	0.58777	0.65014	0.17719
riid	0.08391	0.48145	0.47990	0.11578
rnid	0.08495	0.52742	0.50356	0.12415
wiid	0.15293	0.66096	0.64890	0.16249
wker	0.05503	0.46373	0.56519	0.15932
wnid	0.04752	0.44764	0.55245	0.15663
$\tau = 0.75$				
ksm	0.10563	0.45548	0.44711	0.11185
xy	0.11701	0.61282	0.64894	0.17136
pwy	0.12468	0.63276	0.66372	0.17378
mcmb	0.12674	0.63789	0.67721	0.17864
wxy	0.11557	0.59856	0.62797	0.16437
riid	0.10101	0.50880	0.48791	0.13509
rnid	0.10245	0.51501	0.51436	0.13980
wiid	0.16233	0.70159	0.68879	0.17247
wker	0.06562	0.49482	0.58309	0.16180
wnid	0.08372	0.53969	0.63296	0.17224

Table 9.3: Lengths (L) of the computed 95% confidence intervals of the parameters in the model given by (9.1), for $\tau = 0.50$ and $\tau = 0.75$ quantile functions. Methods used for the estimation are: kernel smooth bootstrap (ksm), (x, y) pair bootstrap (xy); Parzen-Wai-Ying bootstrap (pwy); markov chain marginal bootstrap ($mcmb$); generalised bootstrap method of Bose and Chatterjee (2003) with unit exponential weight (wxy); rank-score test method assuming iid errors ($riid$); rank-score test method assuming nid errors ($rnid$); Wald method assuming iid error using Hall and Sheather's bandwidth ($wiid$); Wald method assuming nid error with Powell's sandwich estimate ($wker$); and Wald method assuming nid error, with Siddiqui sandwicj estimate, using Hall and Sheather's bandwidth ($wnid$).

errors where the transformation to normality is difficult to achieve or maybe undesirable given a need to preserve the original data scale. This gives a great potential for the application of the proposed method.

Bibliography

1. D. W. K. Andrews and M. Buchinsky. A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, 68(1):23–51, 2000.

2. M. Anthony and S. B. Holden. Cross-validation for binary classification by real-valued functions: Theoretical analysis. New York, NY, USA, 1998. Proceedings of the eleventh annual conference on Computational learning theory, COLT 98.

3. G. Bassett and R. Koenker. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363):618–622, 1978.

4. G. Bassett and R. Koenker. An empirical quantile function for linear models with iid errors. *Journal of American Statistical Association*, 77(378):407–415, 1982.

5. E. Bofinger. Estimation of a density function using order statistics. *Australian Journal of Statistics*, 17:1–7, 1975.

6. A. Bose and S. Chatterjee. Generalized bootstrap for estimators of minimizers of convex functions. *Journal of Statistical Planning and Inference*, 117(2):225–239, 2003.

7. A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis.* Oxford University Press, 1997.

8. G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252, 1964.

9. L. D. Brown and M. Levine. Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*, 35(5):2219–2232, 2007.

10. M. Buchinsky. Changes in the u.s. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–458, 1994.

11. M. Buchinsky. Estimating the asymptotic covariance matrix for quantile regression models a monte carlo study. *Journal of Econometrics*, 68(2):303–338, 1995.

12. B. S. Cade, J. D. Richards, and P. W. Mielke Jr. Rank score and permutation testing alternatives for regression quantile estimates. *Journal of Statistical Computation and Simulation*, 76(4):331–355, 2006.

13. E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association,*, 96(455):1022–1030, 2001.

14. V. J. Carey. LMSqreg: An R package for Cole–Green reference centile curves. 2002. URL `http://www.biostat.harvard.edu/~carey`.

15. V. J. Carey, F. H. Youg, L. M. Frenkel, and R. M. McKinney. Growth velocity assessment in pediatric aids: smoothing, penalized quantile regression and the definition of growth failure. *Statistics in Medicine*, 23(3):509–526, 2004.

16. R. J. Carroll. Two examples of transformations when there are possible outliers. *Applied Statistics*, 31:149–152, 1982.

17. Centers for Disease Control and Prevention (CDC). Cut-offs to define outliers in the 2000 CDC Growth Charts. URL `http://www.cdc.gov/nccdphp/dnpa/growthcharts/resources/BIV-cutoffs.pdf`.

18. C. Chen and Y. Wei. Computational issues for quantile regression. *The Indian Journal of Statistics: Special Issue on Quantile Regression and Related Methods*, 67(2):399–417, 2005.

19. C.-K. Chu and J. S. Marron. Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, 19(4):1906–1918, 1991.

20. W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

21. T. J. Cole. Fitting smoothing centile curves to reference data. *Journal of the Royal Statistical Society, Series A–General*, 151(385–418), 1988.

22. T. J. Cole. The lms method for constructing normalized growth standards. *European Journal of Clinical Nutrition*, 44(45–60), 1990.

23. T. J. Cole and P. J. Green. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine*, 11(10):1305–1319, 1992.

24. D. De Angelis, P. Hall, and G. A. Young. Analytical and bootstrap approximations to estimator distribution in l1 regression. *Journal of the American Statistical Association*, 88(424):1310–1316, 1993.

25. P. K. Dunn and G. K. Smyth. Tweedie family densities: Methods of evaluation. In B. Jorgensen, editor, *International Workshop on Statistical Modelling*, In Statistical Modelling, Odense, July 2–6 2001. Proceedings of the 16th International Workshop on Statistical Modelling.

26. P. K. Dunn and G. K. Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15:267–280, 2005.

27. B. Efron. Bootstrap methods: Another look at the jackknife. *Annal of Statistics*, 7(1):1–26, 1979.

28. B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelfia: SIAM, 1982.

29. M. M. El Mouzan, P. J. Foster, A. S. Al Herbish, A. A. Al Salloum, A. A. Al Omar, M. M. Qurachi, and T. Kecojević. The implications of using the world health organization child growth standards in Saudi Arabia. *Nutrition Today*, 44 (2):62–70, 2009.

30. E. Engel. Die productions-und consumptionsverhaeltnisse des koenigsreichs sachsen. *Zeitschrift des Statistischen Bureaus des Koniglich Sachsischen Ministeriums des Inneren*, 8:1–54, 1857.

31. J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, 1996.

32. J. Fan and Q. Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 1998.

33. J. J. Faraway. *Extending the Linear Model with R*. Chapman and Hall, 2006.

34. B. Fitzenberger. The moving block bootstrap and robust inference for linear least square and quantile regression. *Journal of Economics*, 82:235–287, 1997.

35. P. Foster. The local polynomial regression estimator. 2010.

36. J. Fox. *An R and S-Plus companion to applied regression*. SAGE Publications, 2002.

37. J. Fox. Nonparametric regression. 2002. URL `http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf`.

38. T. Gasser, L. Sroka, and C. Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633, 1986.

39. S. Gonçalves and H. White. Bootstrap standard error estimates for linear regression. *Journal of American Statistical Association*, 100(471):970–979, 2005.

40. P. J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55(245–259), 1987.

41. P. J. Green. Penalized likelihood. November 1996.

42. P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman and Hall, 1 edition, 1994.

43. C. Gutenbrunner and J. Jurečková. Regression rank score and regression quantile. *The Annals of Statistics*, 20(1):305–330, 1992.

44. C. Gutenbrunner, J. Jurečková, R. Koenker, and S. Portnoy. Test of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, 2 (4):307–331, 1993.

45. G. Kauermann H. Friedl, A. Berghold, editor. *Leverage Adjustments for Dispersion Modelling in Generalized Nonlinear Models*, Technical University, Graz, Austria, July 19 – 23 1999c. In Statistical Modelling, Proceedings of the 14th International Workshop on Statistical Modelling.

46. J. Hahn. Bootstrapping quantile regression estimators. *Econometric Theory*, 11: 105–121, 1995.

47. J. Hájek and Z. Šidák. *Theory of Rank Test.* New York, Academic Press, 1967.

48. P. Hall and R. J. Carroll. Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society. Series B (Methodological),*, 51(1):3–14, 1989.

49. P. Hall and J. S. Marron. On variance estimation in nonparametric regression. *Biometrika*, 77(2):415–419, 1990.

50. P. Hall and S. Sheather. On the distribution of a studentized quantile. *Journal of the Royal Statistical Society, Series B*, 50:381–391, 1988.

51. P. Hall, J. W. Kay, and D. M. Titterington. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3): 521–528, 1990.

52. X. He and F. Hu. Markov chain marginal bootstrap. *Journal of the American Statistical Association*, 97(459):783–795, 2002.

53. P. J. Huber. Robust estimation of a locatio parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

54. P. J. Huber. *Robust Statistics.* New York John Wiley and Sons, 1981.

55. B. Jorgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):127–162, 1987.

56. M. Kocherginsky, X. He, and Y. Mu. Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14(1):41–55, 2005.

57. R. Koenker. Confidence intervals for regression quantiles. In *Proc. of the 5th Prague Sym. on Asymptotic Stat.*, volume 30, pages 349–359. Springer–Verlag, 1994.

58. R. Koenker. *Quantile Regression.* Cambridge Univerity Press, 2005.

59. R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

60. R. Koenker and G. Bassett. Robast test for heteroscedasticity based on regression quantile. *Econometrica*, 50(1):43–61, 1982.

61. R. Koenker and K. F. Hallock. Quantile regression an introduction. *Journal of Economic Prespectives*, 15(4):143–156, 2001.

62. R. Koenker and J. A. F. Machado. Goodness of fit and relatted inference process for quantile regression. *Journal of the American Statistical Association*, 94(448): 1296–1310, 1999.

63. M. Levine. Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach. *Computational Statistics and Data Analysis*, 50:3405–3431, 2006.

64. M. Levins. *Variance Estimation for Nonparametric Regression and its Application.* PhD thesis, Group of Managerial Science and Applied Economics at the University of Pennsylvania, 2003.

65. C. R. Loader. Bandwidth selection: Classical or plug-in? *The Annals of Statistics*, 27(2):415–438, 1999.

66. C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):661–675, 1973.

67. P. McCullagh and J. A. Nelder. *Generalized Linear Models*. London: Chapman and Hall, 2 edition, 1989.

68. U. U. Müller, A. Schick, and W. Wefelmeyer. Estimating the error variance in nonparametric regression by a covariate-matched u-statistic. *Statistics*, 37: 179–188, 2003.

69. J. A. Nelder and Y. Lee. Generalized linear models for the analysis of taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, 7:107–120, 1991.

70. J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

71. M. H. Neumann. Fully data-driven nonparametric variance estimators. *Statistics*, 25(3):189–212, 1994.

72. M. I. Parzen, L. J. Wei, and Z. Ying. A resampling method based on pivotal estimating function. *Biometrika*, 81(2):341–50, 1994.

73. S. Portnoy. Basic asymptotic. 2010. URL `http://www.stat.uiuc.edu/~sportnoy/chap4.ps`.

74. J. L. Powell. *Estimation of the Monotonic Regression Models under Quantile Restrictions*. Cambridge: Cambridge University Press, 1991.

75. J. Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215–1230, 1984.

76. D. Ruppert and M. P. Wand S. J. Sheather. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90 (432):1257–1270, 1995.

77. A. Sakov and P. J. Bickel. An edgeworth expansion for the m out of n bootstrapped median. *Statistics and Probability Letters*, 49(3):217–223, 2000.

78. M. M. Siddiqui. Distribution of quantiles from a bivariate population. *Journal of Research of the National Bureau of Standards*, 64:145–150, 1960.

79. B. W. Silverman. *Density Estimation*. Chapman and Hall, New York, 1986.

80. B. W. Silverman and G. A. Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479, 1987.

81. G. K. Smyth. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B*, 51(1):47–60, 1989.

82. G. K. Smyth and B. Jorgensen. Fitting tweedie's compound poisson model to insurance claims data: Dispersion modelling. Helsinki, Finland, August 10 – 18 1999. Contributed Paper Meeting 68: Statistics and insurance, In Proceedings of the 52nd Session of the International Statistical Institute.

83. G. K. Smyth and A. P. Verbyla. Double generalized linear models: approximate reml and diagnostics. Technical University, Graz, Austria, July 19 – 23 1999a. Proceedings of the 14th International Workshop on Statistical Modelling, Graz, Austria, Statistical Modelling.

84. G. K. Smyth and A. P. Verbyla. Adjusted likelihood methods for modelling dispersion in general linear models. *Environmetrics*, 10(6):695–709, 1999b.

85. G. K Smyth, A F. Huele, and A. P. Verbyla. Exact and approximate reml for heteroscedastic regression. *Statistical Modelling*, 1:161–175, 2001.

86. M. Stasinopoulos and B. Rigby. Smooth centile curves for skew and kurtotic data modelled using the box-cox power exponential distribution. *Statistics in Medicine*, 23(19):3053–3076, 2004.

87. M. Stasinopoulos, B. Rigby, and C. Akantziliotou. Gamlass: An R package for generalised additive models for location, scale and shape. 2009. URL `http://studweb.north.londonmet.ac.uk/~stasinom/gamlss.html`.

88. J. Taylor. Local polynomial regression and bandwidth selection, February 25 2008. URL `http://maths.dur.ac.uk/Ug/projects/library/CM3/000459698r.pdf`.

89. T. Tong and Y. Wang. Estimating residual variance in nonparametric regression using least squares. *Biometrika*, 92(4):821–830, 2005.

90. J. Tukey. Which part of the sample contains the information? *Proceedings of the National Academy of Science*, 53(1):127–134, 1965.

91. W. N. Venables and B. D. Ripley. Mass: An R package in the standard library of Venables and Ripley. 1998. URL `http://cran.r-project.org/src/contrib/Descriptions/VR.html`.

92. W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. New York: Spriger science+busines media, Inc., 4 edition, 2002.

93. M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

94. R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447, 1974.

95. Y. Wei. *Longitudinal Growth Charts based on semi-parametric qunatile regression*. PhD thesis, University of Illinois at Urbana-Champaign, 2004.

96. Y. Wei and X He. Discussion paper: Conditional growth charts. *The Annals of Statistics*, 35(5):2069–2097, 2006.

97. Y. Wei, A. Pere, R. Koenker, and X. He. Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25:1396–1382, 2006.

98. A. H. Welsh. Asymptotically efficient estimation of the saprsity function at a point. *Statistics and Probability Letters*, 6(6):427–432, 1988.

99. World Helt Organisation WHO. The world health report. 2009.

Appendix A

Wald Test for Girls (0-3) Weight

We express the linear relationship of girls weight on age as:

$$Weight(kg) = \beta_0(\tau) + \beta_1(\tau) \times Age(years).$$

We are interested in the difference between the conditional quantile functions of $\tau_1 = 0.25 = 1/4$ and $\tau_2 = 0.75 = 3/4$. We will consider a general linear hypothesis on the vector $\boldsymbol{\zeta} = \big(\boldsymbol{\beta}(\tau_1)^\top, \boldsymbol{\beta}(\tau_2)^\top\big)$ of the form

$$H_0 \colon \ \boldsymbol{R}\boldsymbol{\zeta} = r,$$

where

$$\boldsymbol{R} = \big(0, -1, 0, 1\big) \quad \text{and} \quad \boldsymbol{\zeta} = \begin{pmatrix} \beta_0(\tau_1) \\ \beta_1(\tau_1) \\ \hline \beta_0(\tau_2) \\ \beta_1(\tau_2) \end{pmatrix},$$

so that

$$\begin{aligned} H_0 &: \beta_1(\tau_1) = \beta_1(\tau_2), \text{ ie. } \beta_1(\tau_2) - \beta_1(\tau_2) = 0 \\ H_1 &: \beta_1(\tau_1) \neq \beta_1(\tau_2). \end{aligned}$$

Using equation (2.21) to calculate the statistic, we need to find $\boldsymbol{R}\big(\boldsymbol{\Omega} \otimes \boldsymbol{Q}^{-1}\big)\boldsymbol{R}^\top$. In our case $\boldsymbol{\Omega}$ is given as

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}$$

where

$$\omega_{11} = \frac{\left(\tau_1 - \tau_1^2\right)}{f^2\left(F^{-1}(\tau_1)\right)}, \quad \omega_{12} = \omega_{21} = \frac{\left(\tau_1 - \tau_1\tau_2\right)}{f\left(F^{-1}(\tau_1)\right) f\left(F^{-1}(\tau_2)\right)} \quad \text{and} \quad \omega_{22} = \frac{\left(\tau_2 - \tau_2^2\right)}{f^2\left(F^{-1}(\tau_2)\right)}.$$

Thus,

$$\boldsymbol{\Omega} \otimes \boldsymbol{Q}^{-1} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} \otimes \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}.$$

Let us addopt the following notation

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Now we can write

$$\boldsymbol{\Omega} \otimes \boldsymbol{Q}^{-1} = \left(\begin{array}{cc|cc} \omega_{11}a & \omega_{11}b & \omega_{12}a & \omega_{12}b \\ \omega_{11}c & \omega_{11}d & \omega_{12}c & \omega_{12}d \\ \hline \omega_{21}a & \omega_{21}b & \omega_{22}a & \omega_{22}b \\ \omega_{21}c & \omega_{21}d & \omega_{22}c & \omega_{22}d \end{array}\right),$$

and find

$$\begin{aligned} \boldsymbol{R}\left(\boldsymbol{\Omega} \otimes \boldsymbol{Q}^{-1}\right)\boldsymbol{R}^\top &= \begin{pmatrix} 0 & -1 & 0 & 1 \end{pmatrix} \left(\begin{array}{cc|cc} \omega_{11}a & \omega_{11}b & \omega_{12}a & \omega_{12}b \\ \omega_{11}c & \omega_{11}d & \omega_{12}c & \omega_{12}d \\ \hline \omega_{21}a & \omega_{21}b & \omega_{22}a & \omega_{22}b \\ \omega_{21}c & \omega_{21}d & \omega_{22}c & \omega_{22}d \end{array}\right) \begin{pmatrix} 0 \\ -1 \\ 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} \omega_{11} & -\omega_{12} & -\omega_{21} & \omega_{22} \end{pmatrix} d \\ &= \begin{pmatrix} \omega_{11} & -2\omega_{12} & \omega_{22} \end{pmatrix} d. \end{aligned}$$

The Hall-Sheather bandwidth for both estimates is 0.03678 (see Figure (2.9) which illustrates the function $\hat{\mathcal{Q}}_Y(\tau|\bar{x})$ for this data set), yielding sparsity estimates of $\hat{s}(0.25) = 6.17831$ and $\hat{s}(0.75) = 5.64860$. Knowing this, we can find

$$\begin{aligned} \omega_{11} &= \left(0.25 - 0.25^2\right) \times 6.17831^2 = 7.15717 \\ \omega_{12} &= \left(0.25 - 0.25 \times 0.75\right) \times 6.17817831 \times 5.64860 = 2.18118 \\ \omega_{22} &= \left(0.75 - 0.75^2\right) \times 5.64860^2 = 5.98251. \end{aligned}$$

Further, we find the the difference in the slopes

$$\hat{\beta}_1(0.75) - \hat{\beta}_1(0.25) = 4.28360 - 3.71779 = 0.56581,$$

and the lower diagonal element of $(X^\top X)^{-1} = 0.00024$. This enables us to calculate the statistic

$$T_n = (0.56581) \times \big((7.15717 - 2 \times 2.18118 + 5.98251) \times 0.00024\big)^{-1} \times (0.56581)\,,$$

and so the test statistic for the equality of the two slopes is 12.42852, which has a p-value less than 0.0001 for one-tailed test of the hypothesis of equality of the slopes.

Appendix B

Resampling to Non-Normal Data

N_iid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.G.Inv_ks	2.00476	0.02051	4.97304	0.03649
glm.G.Inv_ksm	2.00515	0.02042	4.97170	0.03643
glm.G.Inv_N(0, 1)	1.99895	0.02090	4.97276	0.03646
glm.G.Inv_N(0, sd(stu))	1.99959	0.02093	4.97268	0.03653
IRWLS_ks	2.00490	0.02050	4.97140	0.03657
IRWLS_ksm	2.00466	0.02040	4.97136	0.03645
IRWLS_N(0, 1)	1.99962	0.02090	4.97185	0.03650
IRWLS_N(0, sd(stu))	1.99983	0.02092	4.97345	0.03647
Joint_ks	2.00513	0.02040	4.97188	0.03646
Joint_ksm	2.00419	0.02041	4.97313	0.03643
Joint_N(0, 1)	2.00042	0.02089	4.97189	0.03643
Joint_N(0, sd(stu))	2.00003	0.02087	4.97232	0.03644
d_lm_ks	2.00455	0.02045	4.97188	0.03649
d_lm_ksm	2.00534	0.02045	4.97101	0.03651
d_lm_N(0, 1)	1.99904	0.02088	4.97411	0.03652
d_lm_N(0, sd(stu))	2.00158	0.02090	4.96956	0.03648
d_rlm_ks	2.00447	0.02045	4.97359	0.03646
d_rlm_ksm	2.00514	0.02041	4.97056	0.03656
d_rlm_N(0, 1)	2.00009	0.02096	4.97185	0.03652
d_rlm_N(0, sd(stu))	1.99938	0.02090	4.97334	0.03642
d_lws_ks	2.00566	0.02048	4.97022	0.03650
d_lws_ksm	2.00435	0.02042	4.97345	0.03648
d_lws_N(0, 1)	1.99978	0.02092	4.97252	0.03650
d_lws_N(0, sd(stu))	2.00001	0.02092	4.97194	0.03648
glm.G.log_ks	2.00676	0.02045	4.96841	0.03650
glm.G.log_ksm	2.00469	0.02046	4.97217	0.03654
glm.G.log_N(0, 1)	2.00065	0.02095	4.97233	0.03655
glm.G.log_N(0, sd(stu))	2.00010	0.02089	4.97134	0.03643
xy	2.00254	0.02021	4.97388	0.03508
pwy	2.00293	0.02017	4.97237	0.03496
mcmb	2.00082	0.02017	4.97622	0.03514
wxy	2.00359	0.02017	4.97161	0.03499
riid	2.00033	0.02091	4.97187	0.03647
rnid	2.00033	0.02091	4.97187	0.03647
wiid	2.00033	0.02091	4.97187	0.03647
wker	2.00033	0.02091	4.97187	0.03647
wnid	2.00033	0.02091	4.97187	0.03647

Table B.1: Parameter estimates for Model 1 for $\tau = 0.5$.

N_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.G.Inv_ks	95.60%	1.82247	0.00487	95.20%	3.15669	0.00829
glm.G.Inv_ksm	94.40%	1.74229	0.00461	94.60%	3.01729	0.00818
glm.G.Inv_N(0, 1)	93.80%	1.75126	0.00313	94.80%	3.03064	0.00542
glm.G.Inv_N(0, sd(stu))	94.20%	1.75181	0.00303	94.80%	3.03684	0.00533
IRWLS_ks	95.00%	1.82206	0.00480	94.80%	3.15420	0.00828
IRWLS_ksm	94.40%	1.74259	0.00468	93.60%	3.01627	0.00795
IRWLS_N(0, 1)	94.00%	1.74341	0.00390	94.20%	3.02920	0.00557
IRWLS_N(0, sd(stu))	94.00%	1.74550	0.00379	94.40%	3.03280	0.00539
Joint_ks	95.00%	1.82317	0.00475	95.00%	3.15876	0.00830
Joint_ksm	94.20%	1.74361	0.00457	94.60%	3.02153	0.00794
Joint_N(0, 1)	94.00%	1.74448	0.00386	94.40%	3.03027	0.00558
Joint_N(0, sd(stu))	94.00%	1.74694	0.00395	94.40%	3.03487	0.00544
d_lm_ks	95.20%	1.82447	0.00499	95.40%	3.15792	0.00839
d_lm_ksm	94.60%	1.74415	0.00472	94.80%	3.02561	0.00814
d_lm_N(0, 1)	95.80%	1.97175	0.00452	96.20%	3.41345	0.00674
d_lm_N(0, sd(stu))	94.00%	1.75350	0.00384	94.80%	3.03998	0.00541
d_rlm_ks	95.00%	1.82898	0.00481	95.00%	3.16816	0.00840
d_rlm_ksm	94.40%	1.74615	0.00462	94.20%	3.02154	0.00791
d_rlm_N(0, 1)	93.80%	1.72688	0.00396	93.60%	2.99493	0.00643
d_rlm_N(0, sd(stu))	93.60%	1.75271	0.00366	94.40%	3.03841	0.00555
d_lws_ks	95.00%	1.82750	0.00479	95.00%	3.16693	0.00820
d_lws_ksm	94.40%	1.74736	0.00456	94.60%	3.02609	0.00790
d_lws_N(0, 1)	91.20%	1.54965	0.00439	91.20%	2.68861	0.00726
d_lws_N(0, sd(stu))	94.00%	1.75153	0.00403	94.40%	3.04130	0.00601
glm.G.log_ks	94.80%	1.82260	0.00484	95.00%	3.16115	0.00845
glm.G.log_ksm	94.40%	1.74271	0.00453	93.80%	3.01690	0.00775
glm.G.log_N(0, 1)	94.20%	1.75098	0.00331	94.20%	3.03360	0.00574
glm.G.log_N(0, sd(stu))	94.00%	1.75405	0.00311	94.20%	3.04071	0.00531
xy	94.60%	1.79069	0.01465	94.20%	3.12683	0.02283
pwy	93.80%	1.79675	0.01465	94.00%	3.13368	0.02278
mcmb	94.40%	1.77754	0.01228	94.20%	3.08393	0.02210
wxy	94.20%	1.79034	0.01462	93.80%	3.12336	0.02297
riid	88.80%	1.45275	0.01219	91.80%	2.53557	0.01943
rnid	89.00%	1.45394	0.01219	91.80%	2.54626	0.01958
wiid	93.40%	1.74107	0.01082	93.60%	3.01412	0.01874
wker	96.00%	2.04935	0.00485	96.60%	3.54991	0.00796
wnid	94.00%	1.76758	0.00830	95.00%	3.06231	0.01093

Table B.2: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.Inv	1.00161	0.00004
IRWLS	1.00161	0.00004
Joint Modelling	1.00162	0.00004
d(adj)_lm	0.88982	0.00092
d(adj)_rlm	1.01517	0.00136
d(adj)_lowess	1.13277	0.00212
glm.Gamma.log	1.00161	0.00004

Table B.3: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) conditional variance function estimate.

N_nid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.G.Inv_ks	2.00742	0.02734	4.95032	0.05913
glm.G.Inv_ksm	2.00819	0.02722	4.94793	0.05906
glm.G.Inv_N(0, 1)	2.00087	0.02779	4.94902	0.05914
glm.G.Inv_N(0, sd(stu))	2.00147	0.02782	4.94972	0.05921
IRWLS_ks	2.00770	0.02731	4.94766	0.05927
IRWLS_ksm	2.00736	0.02719	4.94758	0.05910
IRWLS_N(0, 1)	2.00178	0.02778	4.94756	0.05920
IRWLS_N(0, sd(stu))	2.00178	0.02781	4.95122	0.05910
Joint_ks	2.00833	0.02720	4.94809	0.05909
Joint_ksm	2.00649	0.02724	4.95036	0.05904
Joint_N(0, 1)	2.00252	0.02776	4.94894	0.05906
Joint_N(0, sd(stu))	2.00185	0.02774	4.94972	0.05908
d_lm_ks	2.00673	0.02725	4.94900	0.05912
d_lm_ksm	2.00836	0.02727	4.94677	0.05911
d_lm_N(0, 1)	2.00062	0.02776	4.95217	0.05923
d_lm_N(0, sd(stu))	2.00441	0.02778	4.94479	0.05914
d_rlm_ks	2.00656	0.02725	4.95178	0.05911
d_rlm_ksm	2.00798	0.02722	4.94639	0.05928
d_rlm_N(0, 1)	2.00221	0.02786	4.94832	0.05916
d_rlm_N(0, sd(stu))	2.00131	0.02778	4.95062	0.05900
d_lws_ks	2.00875	0.02730	4.94588	0.05915
d_lws_ksm	2.00628	0.02722	4.95133	0.0510
d_lws_N(0, 1)	2.00177	0.02781	4.94937	0.05920
d_lws_N(0, sd(stu))	2.00202	0.02781	4.94864	0.05908
glm.G.log_ks	2.01052	0.02729	4.94305	0.05912
glm.G.log_ksm	2.00697	0.02734	4.94935	0.05922
glm.G.log_N(0, 1)	2.00295	0.02784	4.94975	0.05923
glm.G.log_N(0, sd(stu))	2.00233	0.02777	4.94727	0.05906
xy	2.00699	0.02679	4.94454	0.05694
pwy	2.00733	0.02670	4.94222	0.05677
mcmb	2.00411	0.02682	4.94902	0.05711
wxy	2.00831	0.02672	4.94115	0.05682
riid	2.00251	0.02779	4.94860	0.05911
rnid	2.00251	0.02779	4.94860	0.05911
wiid	2.00251	0.02779	4.94860	0.05911
wker	2.00251	0.02779	4.94860	0.05911
wnid	2.00251	0.02779	4.94860	0.05911

Table B.4: Parameter estimates for Model 1 for $\tau = 0.5$.

N_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.G.Inv_ks	98.20%	2.95598	0.00787	95.20%	5.11927	0.01333
glm.G.Inv_ksm	97.60%	2.82533	0.00746	94.20%	4.89381	0.01328
glm.G.Inv_N(0, 1)	95.60%	2.54465	0.00493	95.80%	5.29510	0.00985
glm.G.Inv_N(0, sd(stu))	95.80%	2.54604	0.00480	95.80%	5.30666	0.01004
IRWLS_ks	97.80%	2.95749	0.00778	95.00%	5.11915	0.01349
IRWLS_ksm	97.40%	2.82546	0.00757	94.20%	4.89095	0.01280
IRWLS_N(0, 1)	95.60%	2.54043	0.00509	96.00%	5.28949	0.01016
IRWLS_N(0, sd(stu))	95.60%	2.54281	0.00486	95.40%	5.29746	0.00984
Joint_ks	98.00%	2.95042	0.00785	94.40%	5.11320	0.01369
Joint_ksm	97.60%	2.82216	0.00753	93.80%	4.89020	0.01302
Joint_N(0, 1)	94.00%	2.31298	0.00592	94.00%	4.96810	0.00917
Joint_N(0, sd(stu))	94.20%	2.31665	0.00596	94.20%	4.97755	0.00895
d_lm_ks	98.40%	2.96010	0.00800	95.20%	5.12289	0.01356
d_lm_ksm	97.80%	2.82985	0.00758	94.20%	4.90810	0.01306
d_lm_N(0, 1)	97.40%	2.86634	0.00689	97.40%	5.94129	0.01250
d_lm_N(0, sd(stu))	95.60%	2.55737	0.00580	96.20%	5.30830	0.00989
d_rlm_ks	98.20%	2.96524	0.00784	95.20%	5.13721	0.01366
d_rlm_ksm	97.80%	2.83277	0.00753	93.60%	4.90292	0.01284
d_rlm_N(0, 1)	95.40%	2.51420	0.00606	96.00%	5.22422	0.01179
d_rlm_N(0, sd(stu))	95.80%	2.55375	0.00571	95.80%	5.30071	0.01018
d_lws_ks	98.80%	2.96340	0.00777	95.20%	5.13590	0.01331
d_lws_ksm	97.60%	2.83480	0.00737	94.00%	4.91120	0.01280
d_lws_N(0, 1)	93.40%	2.26267	0.00670	92.20%	4.70489	0.01299
d_lws_N(0, sd(stu))	95.60%	2.55779	0.00650	96.00%	5.32260	0.01105
glm.G.log_ks	98.60%	2.95089	0.00793	95.20%	5.11802	0.01380
glm.G.log_ksm	97.60%	2.82218	0.00743	93.80%	4.88488	0.01272
glm.G.log_N(0, 1)	94.60%	2.41110	0.00544	95.80%	5.17763	0.00995
glm.G.log_N(0, sd(stu))	94.60%	2.41508	0.00525	95.40%	5.18791	0.00928
xy	94.20%	2.38290	0.01944	93.80%	5.12862	0.03706
pwy	94.00%	2.39051	0.01943	94.20%	5.13710	0.03683
mcmb	96.40%	2.59691	0.01742	95.20%	5.18707	0.03791
wxy	93.80%	2.38179	0.01939	94.20%	5.12363	0.03745
riid	88.80%	1.87258	0.01583	88.40%	3.96739	0.03280
rnid	88.80%	1.87258	0.01583	88.40%	3.96739	0.03280
wiid	96.60%	2.81890	0.01744	93.40%	4.88004	0.03019
wker	98.00%	2.96329	0.00638	97.40%	5.91762	0.01438
wnid	93.60%	2.38726	0.01226	95.60%	5.11895	0.01833

Table B.5: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.Inv	1.00140	0.00003
IRWLS	1.00140	0.00003
Joint Modelling	1.00210	0.00023
d(adj)_lm	0.89212	0.00091
d(adj)_rlm	1.01623	0.00132
d(adj)_lowess	1.13263	0.00207
glm.Gamma.log	1.00141	0.00004

Table B.6: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) conditional variance function estimate.

t(20)_iid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.G.Inv_ks	1.99152	0.00488	5.01418	0.00855
glm.G.Inv_ksm	1.99155	0.00487	5.01434	0.00856
glm.G.Inv_t(20)	1.99026	0.00495	5.01482	0.00857
IRWLS_ks	1.99185	0.00488	5.01353	0.00856
IRWLS_ksm	1.99172	0.00487	5.01410	0.00857
IRWLS_t(20)	1.99041	0.00494	5.01478	0.00856
Joint_ks	1.99129	0.00489	5.01444	0.00858
Joint_ksm	1.99140	0.00487	5.01464	0.00855
Joint_t(20)	1.99047	0.00495	5.01431	0.00856
d_lm_ks	1.99130	0.00490	5.01448	0.00858
d_lm_ksm	1.99177	0.00487	5.01411	0.00857
d_lm_t(20)	1.99046	0.00496	5.01444	0.00857
d_rlm_ks	1.99152	0.00487	5.01418	0.00853
d_rlm_ksm	1.99140	0.00487	5.01454	0.00856
d_rlm_t(20)	1.99043	0.00494	5.01453	0.00857
d_lws_ks	1.99150	0.00489	5.01434	0.00856
d_lws_ksm	1.99165	0.00489	5.01428	0.00858
d_lws_t(20)	1.99038	0.00495	5.01434	0.00855
glm.G.log_ks	1.99158	0.00487	5.01416	0.00855
glm.G.log_ksm	1.99129	0.00487	5.01500	0.00856
glm.G.log_t(20)	1.99018	0.00496	5.01461	0.00857
xy	1.99166	0.00482	5.01299	0.00828
pwy	1.99139	0.00483	5.01315	0.00829
mcmb	1.99082	0.00484	5.01425	0.00834
wxy	1.99151	0.00482	5.01314	0.00830
riid	1.99041	0.00495	5.01447	0.00856
rnid	1.99041	0.00495	5.01447	0.00856
wiid	1.99041	0.00495	5.01447	0.00856
wker	1.99041	0.00495	5.01447	0.00856
wnid	1.99041	0.00495	5.01447	0.00856

Table B.7: Parameter estimates for Model 1 for $\tau = 0.5$.

t(20)_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.G.Inv_ks	96.80%	0.46421	0.00117	95.00%	0.80450	0.00204
glm.G.Inv_ksm	95.20%	0.44461	0.00115	93.60%	0.76951	0.00195
glm.G.Inv_t(20)	96.20%	0.46635	0.00086	95.80%	0.80800	0.00152
IRWLS_ks	96.60%	0.46526	0.00121	95.20%	0.80559	0.00212
IRWLS_ksm	95.40%	0.44491	0.00119	93.80%	0.76965	0.00204
IRWLS_t(20)	96.60%	0.46687	0.00114	95.20%	0.80670	0.00153
Joint_ks	96.40%	0.46479	0.00123	94.80%	0.80493	0.00209
Joint_ksm	95.20%	0.44443	0.00114	94.40%	0.77073	0.00196
Joint_t(20)	97.20%	0.46550	0.00116	95.60%	0.80540	0.00153
d_lm_ks	96.60%	0.46521	0.00119	95.00%	0.80623	0.00203
d_lm_ksm	96.00%	0.44556	0.00114	93.80%	0.77187	0.00195
d_lm_t(20)	98.00%	0.51766	0.00121	97.60%	0.89695	0.00184
d_rlm_ks	96.40%	0.46616	0.00116	95.00%	0.80689	0.00201
d_rlm_ksm	95.00%	0.44455	0.00117	94.00%	0.76916	0.00199
d_rlm_t(20)	95.40%	0.45280	0.00106	94.00%	0.78485	0.00172
d_lws_ks	96.40%	0.46624	0.00123	95.40%	0.80725	0.00207
d_lws_ksm	95.40%	0.44503	0.00111	93.60%	0.77111	0.00194
d_lws_t(20)	93.60%	0.40615	0.00118	92.40%	0.70530	0.00196
glm.G.log_ks	96.40%	0.46546	0.00119	95.00%	0.80553	0.00207
glm.G.log_ksm	95.40%	0.44417	0.00114	94.00%	0.77009	0.00197
glm.G.log_t(20)	95.60%	0.46645	0.00089	95.40%	0.80766	0.00153
xy	95.60%	0.45506	0.00338	94.60%	0.79059	0.00506
pwy	96.20%	0.45638	0.00345	94.60%	0.79362	0.00515
mcmb	95.60%	0.45190	0.00287	94.80%	0.78230	0.00498
wxy	96.20%	0.45561	0.00339	94.40%	0.79045	0.00511
riid	88.80%	0.36348	0.00329	90.80%	0.63475	0.00521
rnid	88.80%	0.36382	0.00329	91.00%	0.63702	0.00524
wiid	95.40%	0.44454	0.00278	92.40%	0.76958	0.00481
wker	98.40%	0.52671	0.00130	98.00%	0.91044	0.00209
wnid	95.20%	0.45155	0.00214	94.20%	0.77967	0.00282

Table B.8: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.Inv	1.00156	0.00004
IRWLS	1.00155	0.00004
Joint Modelling	1.00156	0.00005
d(adj)_lm	0.90275	0.00087
d(adj)_rlm	1.03019	0.00130
d(adj)_lowess	1.14882	0.00213
glm.Gamma.log	1.00156	0.00004

Table B.9: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) conditional variance function estimate.

t(20)_nid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.G.Inv_ks	1.99708	0.00633	5.00324	0.01435
glm.G.Inv_ksm	1.99704	0.00631	5.00292	0.01434
glm.G.Inv_t(20)	1.99907	0.00650	5.00279	0.01430
IRWLS_ks	1.99710	0.00632	5.00326	0.01430
IRWLS_ksm	1.99682	0.00633	5.00351	0.01434
IRWLS_t(20)	1.99898	0.00649	5.00317	0.01432
Joint_ks	1.99735	0.00634	5.00258	0.01432
Joint_ksm	1.99701	0.00632	5.00315	0.01433
Joint_t(20)	1.99967	0.00648	5.00230	0.01433
d_lm_ks	1.99725	0.0063	5.00298	0.01432
d_lm_ksm	1.99733	0.00632	5.00280	0.01430
d_lm_t(20)	1.99934	0.00648	5.00283	0.01434
d_rlm_ks	1.99711	0.00632	5.00280	0.01430
d_rlm_ksm	1.99703	0.00633	5.00274	0.01432
d_rlm_t(20)	1.99879	0.00648	5.00349	0.01431
d_lws_ks	1.99717	0.00631	5.00309	0.01429
d_lws_ksm	1.99714	0.00632	5.00339	0.01432
d_lws_t(20)	1.99946	0.00650	5.00216	0.01431
glm.G.log_ks	1.99729	0.00631	5.00278	0.01431
glm.G.log_ksm	1.99712	0.00632	5.00307	0.01432
glm.G.log_t(20)	1.99882	0.00648	5.00371	0.01434
xy	2.00056	0.00611	4.99794	0.01349
pwy	2.00065	0.00611	4.99808	0.01350
mcmb	1.99994	0.00617	4.99964	0.01363
wxy	2.00028	0.00610	4.99822	0.01350
riid	1.99920	0.00648	5.00305	0.01431
rnid	1.99920	0.00648	5.00305	0.014301
wiid	1.99920	0.00648	5.00305	0.01431
wker	1.99920	0.00648	5.00305	0.01431
wnid	1.99920	0.00648	5.00305	0.01431

Table B.10: Parameter estimates for Model 1 for $\tau = 0.5$.

t(20)_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.G.Inv_ks	99.40%	0.75563	0.00204	95.20%	1.30803	0.00349
glm.G.Inv_ksm	99.00%	0.72238	0.00188	94.60%	1.24990	0.00332
glm.G.Inv_t(20)	98.00%	0.67790	0.00136	97.00%	1.40550	0.00272
IRWLS_ks	99.20%	0.75524	0.00197	96.00%	1.30801	0.00336
IRWLS_ksm	99.00%	0.72190	0.00192	94.40%	1.25010	0.00334
IRWLS_t(20)	98.20%	0.67893	0.00134	96.80%	1.40580	0.00268
Joint_ks	99.00%	0.75267	0.00200	95.60%	1.30445	0.00348
Joint_ksm	99.00%	0.72160	0.00189	94.60%	1.24933	0.00332
Joint_t(20)	97.40%	0.61978	0.00158	96.20%	1.32249	0.00253
d_lm_ks	99.20%	0.75620	0.00198	95.40%	1.30887	0.00346
d_lm_ksm	99.20%	0.72384	0.00190	94.40%	1.25330	0.00329
d_lm_t(20)	99.20%	0.75799	0.00186	98.80%	1.55596	0.00328
d_rlm_ks	99.40%	0.75631	0.00201	95.80%	1.30916	0.00343
d_rlm_ksm	99.20%	0.72429	0.00193	94.80%	1.25491	0.00333
d_rlm_t(20)	98.40%	0.66680	0.00159	96.20%	1.37092	0.00324
d_lws_ks	99.40%	0.75915	0.00199	96.60%	1.31352	0.00343
d_lws_ksm	99.20%	0.72528	0.00192	95.00%	1.25573	0.00336
d_lws_t(20)	97.80%	0.60255	0.00172	94.80%	1.23892	0.00358
glm.G.log_ks	99.20%	0.75486	0.00206	95.60%	1.30783	0.00352
glm.G.log_ksm	99.20%	0.72232	0.00192	94.80%	1.25026	0.00331
glm.G.log_t(20)	97.80%	0.64549	0.00142	96.80%	1.37621	0.00258
xy	97.40%	0.60468	0.00479	96.40%	1.30499	0.00914
pwy	97.20%	0.60592	0.00475	96.40%	1.30852	0.00915
mcmb	98.80%	0.66416	0.00441	96.80%	1.33040	0.00951
wxy	97.40%	0.60443	0.00479	96.40%	1.30366	0.00916
riid	90.60%	0.48274	0.00430	89.80%	1.00976	0.00856
rnid	90.60%	0.48596	0.00426	90.60%	1.03774	0.00866
wiid	99.00%	0.71753	0.00447	94.40%	1.24217	0.00774
wker	99.60%	0.76321	0.00165	98.40%	1.52229	0.00366
wnid	96.60%	0.60413	0.00323	95.20%	1.30799	0.00475

Table B.11: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.Inv	1.00138	0.00003
IRWLS	1.00138	0.00003
Joint Modelling	1.00162	0.00023
d(adj)_lm	0.90373	0.00092
d(adj)_rlm	1.02730	0.00129
d(adj)_lowess	1.14193	0.00196
glm.Gamma.log	1.00139	0.00004

Table B.12: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) conditional variance function estimate.

t(10)_iid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.G.Inv_ks	2.00044	0.00486	4.99432	0.00887
glm.G.Inv_ksm	2.00048	0.00485	4.99426	0.00886
glm.G.Inv_t(10)	2.00006	0.00503	4.99410	0.00889
IRWLS_ks	2.00061	0.00482	4.99418	0.00882
IRWLS_ksm	2.00045	0.00485	4.99407	0.00886
IRWLS_t(10)	2.00004	0.00501	4.99419	0.00887
Joint_ks	2.00047	0.00486	4.99423	0.00886
Joint_ksm	2.00020	0.00485	4.99475	0.00886
Joint_t(10)	2.00024	0.00502	4.99390	0.00885
d_lm_ks	2.00032	0.00486	4.99440	0.00886
d_lm_ksm	2.00036	0.00484	4.99431	0.00884
d_lm_t(10)	1.99977	0.00502	4.99437	0.00888
d_rlm_ks	2.00060	0.00484	4.99405	0.00885
d_rlm_ksm	2.00046	0.00486	4.99439	0.00887
d_rlm_t(10)	2.00004	0.00501	4.99390	0.00885
d_lws_ks	2.00048	0.00486	4.99426	0.00887
d_lws_ksm	2.00063	0.00484	4.99435	0.00886
d_lws_t(10)	1.99999	0.00502	4.99426	0.00887
glm.G.log_ks	2.00048	0.00485	4.99415	0.00889
glm.G.log_ksm	2.00041	0.00485	4.99431	0.00887
glm.G.log_t(10)	2.00001	0.00502	4.99438	0.00886
xy	1.99880	0.00477	4.99648	0.00854
pwy	1.99890	0.00477	4.99650	0.00853
mcmb	1.99906	0.00476	4.99608	0.00854
wxy	1.99880	0.00477	4.99652	0.00854
riid	2.00000	0.00501	4.99430	0.00885
rnid	2.00000	0.00501	4.99430	0.00885
wiid	2.00000	0.00501	4.99430	0.00885
wker	2.00000	0.00501	4.99430	0.00885
wnid	2.00000	0.00501	4.99430	0.00885

Table B.13: Parameter estimates for Model 1 for $\tau = 0.5$.

t(10)_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.G.Inv_ks	96.40%	0.47640	0.00122	95.00%	0.82498	0.00212
glm.G.Inv_ksm	95.20%	0.45583	0.00116	93.60%	0.78861	0.00203
glm.G.Inv_t(20)	97.20%	0.49955	0.00097	96.80%	0.86563	0.00166
IRWLS_ks	96.60%	0.47660	0.00122	94.80%	0.82576	0.00209
IRWLS_ksm	95.80%	0.45496	0.00114	93.40%	0.78814	0.00199
IRWLS_t(10)	96.80%	0.49979	0.00120	96.40%	0.86473	0.00170
Joint_ks	96.40%	0.47600	0.00121	95.20%	0.82515	0.00210
Joint_ksm	95.60%	0.45464	0.00117	93.20%	0.78879	0.00208
Joint_t(10)	97.00%	0.49901	0.00123	96.80%	0.86278	0.00162
d_lm_ks	96.60%	0.47666	0.00119	94.60%	0.82608	0.00207
d_lm_ksm	95.60%	0.45594	0.00115	93.00%	0.78892	0.00202
d_lm_t(10)	97.80%	0.54633	0.00131	98.00%	0.94703	0.00194
d_rlm_ks	97.40%	0.47747	0.00123	95.60%	0.82572	0.00214
d_rlm_ksm	95.40%	0.45629	0.00119	93.40%	0.78998	0.00206
d_rlm_t(10)	96.60%	0.48167	0.00114	95.20%	0.83414	0.00179
d_lws_ks	96.80%	0.47814	0.00123	95.20%	0.82792	0.00210
d_lws_ksm	95.80%	0.45712	0.00115	93.00%	0.79149	0.00197
d_lws_t(10)	94.00%	0.43388	0.00122	92.40%	0.75155	0.00204
glm.G.log_ks	96.80%	0.47663	0.00121	95.00%	0.82517	0.00210
glm.G.log_ksm	95.60%	0.45474	0.00111	94.00%	0.78835	0.00194
glm.G.log_t(10)	97.00%	0.49934	0.00094	96.80%	0.86419	0.00163
xy	95.60%	0.46613	0.00377	93.80%	0.80932	0.00562
pwy	95.80%	0.46635	0.00372	94.00%	0.81039	0.00561
mcmb	94.80%	0.46205	0.00316	94.20%	0.79874	0.00535
wxy	95.60%	0.46536	0.00371	94.20%	0.80793	0.00553
riid	89.40%	0.37257	0.00305	90.00%	0.64352	0.00519
rnid	89.40%	0.37289	0.00305	90.00%	0.64600	0.00518
wiid	94.00%	0.45028	0.00282	92.20%	0.77951	0.00489
wker	97.60%	0.54069	0.00141	98.20%	0.93651	0.00227
wnid	94.40%	0.45849	0.00221	93.40%	0.79282	0.00287

Table B.14: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.Inv	1.00158	0.00004
IRWLS	1.00157	0.00004
Joint Modelling	1.00156	0.00005
d(adj)_lm	0.91547	0.00089
d(adj)_rlm	1.04050	0.00123
d(adj)_lowess	1.15666	0.00195
glm.Gamma.log	1.00158	0.00004

Table B.15: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) conditional variance function estimate.

t(10)_nid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.G.Inv_ks	1.99378	0.00668	5.02282	0.01420
glm.G.Inv_ksm	1.99327	0.00668	5.02358	0.01421
glm.G.Inv_t(10)	1.99280	0.00681	5.02277	0.01420
IRWLS_ks	1.99331	0.00671	5.02370	0.01421
IRWLS_ksm	1.99313	0.00667	5.02397	0.01418
IRWLS_t(10)	1.99279	0.00681	5.02276	0.01420
Joint_ks	1.99404	0.00668	5.02289	0.01420
Joint_ksm	1.99374	0.00668	5.02293	0.01419
Joint_t(10)	1.99281	0.00682	5.02277	0.01420
d_lm_ks	1.99374	0.00668	5.02310	0.01419
d_lm_ksm	1.99346	0.00669	5.02302	0.01420
d_lm_t(10)	1.99268	0.00683	5.02320	0.01420
d_rlm_ks	1.99358	0.00669	5.02316	0.01420
d_rlm_ksm	1.99357	0.00668	5.02323	0.01419
d_rlm_t(10)	1.99302	0.00681	5.02224	0.01421
d_lws_ks	1.99317	0.00670	5.02392	0.01422
d_lws_ksm	1.99343	0.00667	5.02345	0.01416
d_lws_t(10)	1.99235	0.00684	5.02356	0.0143
glm.G.log_ks	1.99341	0.00666	5.02343	0.01416
glm.G.log_ksm	1.99365	0.00667	5.02284	0.01418
glm.G.log_T(10)	1.99253	0.00682	5.02311	0.01426
xy	1.99540	0.00657	5.01658	0.01375
pwy	1.99576	0.00658	5.01580	0.01382
mcmb	1.99467	0.00655	5.01931	0.01377
wxy	1.99567	0.00656	5.01671	0.01377
riid	1.99273	0.00682	5.02314	0.01419
rnid	1.99273	0.00682	5.02314	0.01419
wiid	1.99273	0.00682	5.02314	0.01419
wker	1.99273	0.00682	5.02314	0.01419
wnid	1.99273	0.00682	5.02314	0.01419

Table B.16: Parameter estimates for Model 1 for $\tau = 0.5$.

t(10)_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.G.Inv_ks	99.20%	0.77788	0.00204	96.40%	1.34589	0.00354
glm.G.Inv_ksm	98.80%	0.74197	0.00188	95.40%	1.28348	0.00328
glm.G.Inv_t(0)	98.40%	0.72799	0.00150	97.60%	1.50052	0.00303
IRWLS_ks	98.60%	0.77610	0.00198	96.60%	1.34532	0.00340
IRWLS_ksm	98.80%	0.74182	0.00192	95.00%	1.28447	0.00330
IRWLS_t(0)	98.40%	0.72674	0.00146	97.80%	1.50005	0.00304
Joint_ks	98.60%	0.77418	0.00206	96.60%	1.33997	0.00355
Joint_ksm	98.20%	0.74120	0.00194	95.80%	1.28339	0.00330
Joint_t(20)	97.40%	0.66752	0.00185	97.20%	1.41565	0.00272
d_lm_ks	99.00%	0.77830	0.00202	96.80%	1.34952	0.00351
d_lm_ksm	98.00%	0.74255	0.00187	95.40%	1.28661	0.00326
d_lm_t(20)	99.20%	0.80253	0.00208	98.60%	1.63921	0.00369
d_rlm_ks	98.80%	0.77894	0.00204	96.60%	1.34795	0.00350
d_rlm_ksm	98.60%	0.74443	0.00190	95.20%	1.28924	0.00329
d_rlm_t(20)	98.00%	0.70478	0.00176	97.60%	1.44060	0.00338
d_lws_ks	99.00%	0.77802	0.00208	96.60%	1.34737	0.00353
d_lws_ksm	98.40%	0.74387	0.00187	95.80%	1.28849	0.00322
d_lws_t(20)	96.40%	0.64049	0.00201	96.20%	1.30908	0.00394
glm.G.log_ks	98.80%	0.77409	0.00203	96.80%	1.34171	0.00355
glm.G.log_ksm	98.40%	0.74192	0.00192	95.40%	1.28396	0.00336
glm.G.log_t(20)	98.20%	0.69240	0.00168	97.40%	1.46661	0.00297
xy	95.40%	0.61804	0.00503	95.80%	1.32238	0.00917
pwy	95.60%	0.62058	0.00501	96.00%	1.32671	0.00914
mcmb	96.60%	0.67346	0.00445	96.40%	1.34145	0.00938
wxy	95.40%	0.61786	0.00504	95.80%	1.31996	0.00914
riid	90.80%	0.48751	0.00394	89.90%	1.03329	0.00891
rnid	91.00%	0.49104	0.00392	90.20%	1.06312	0.00909
wiid	97.00%	0.73013	0.00450	95.00%	1.26399	0.00780
wker	99.00%	0.78653	0.00179	98.20%	1.56268	0.00398
wnid	95.40%	0.61979	0.00317	96.80%	1.33256	0.00488

Table B.17: Results for Model 2 with $e_i \sim t(10)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.Inv	1.00142	0.00003
IRWLS	1.00142	0.00003
Joint Modelling	1.00151	0.00022
d(adj)_lm	0.91558	0.00090
d(adj)_rlm	1.04064	0.00131
d(adj)_lowess	1.15635	0.00208
glm.Gamma.log	1.00143	0.00004

Table B.18: Results for Model 2 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Iteratively Reweighted Least Squares, Joint Modelling, Dispersion Adjusted (using lm, rlm and lowess) and gamma GLM (link=log) conditional variance function estimate.

Appendix C

Extending GLMs for Resampling Schemes

N_iid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.98074	0.01880	5.03452	0.03440
glm.Tw.0_N(0, 1)	1.97959	0.01943	5.03402	0.03442
dglm.Tw.0_ksm	1.97980	0.01878	5.03539	0.03436
dglm.Tw.0_N(0, 1)	1.97814	0.01942	5.03548	0.03443
glmT0-dglmT0_ksm	1.97967	0.01877	5.03494	0.03438
glmT0-dglmT0_N(0, 1)	1.97900	0.01946	5.03493	0.03449
robglm.G.log_ksm	1.97860	0.01883	5.03587	0.03447
robglm.G.log_N(0, 1)	1.97747	0.01943	5.03662	0.03441
xy	1.98680	0.01888	5.02141	0.03344
pwy	1.98866	0.01886	5.01743	0.03338
mcmb	1.98660	0.01881	5.02244	0.03339
wxy	1.98834	0.01891	5.01982	0.03340
riid	1.97834	0.01941	5.03588	0.03441
rnid	1.97834	0.01941	5.03588	0.03441
wiid	1.97834	0.01941	5.03588	0.03441
wker	1.97834	0.01941	5.03588	0.03441
wnid	1.97834	0.01941	5.03588	0.03441

Table C.1: Parameter estimates for Model 1 for $\tau = 0.5$.

N_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	95.80%	1.75012	0.00466	95.40%	3.03077	0.00805
glm.Tw.0_N(0, 1)	95.20%	1.75116	0.00306	96.20%	3.03221	0.00523
dglm.Tw.0_ksm	95.60%	1.74894	0.00479	95.40%	3.02712	0.00844
dglm.Tw.0_N(0, 1)	95.60%	1.74971	0.00375	96.20%	3.03410	0.00540
glmT0-dglmT0_ksm	95.60%	1.75324	0.00475	95.80%	3.03411	0.00814
glmT0-dglmT0_N(0, 1)	90.60%	1.45353	0.00327	88.80%	2.51613	0.00493
robglm.G.log_ksm	95.80%	1.74784	0.00481	95.40%	3.02773	0.00826
robglm.G.log_N(0, 1)	95.40%	1.75121	0.00340	96.00%	3.03079	0.00582
xy	95.00%	1.80082	0.01448	95.00%	3.12648	0.02180
pwy	94.80%	1.80636	0.01440	94.80%	3.13854	0.02204
mcmb	95.40%	1.79065	0.01239	94.80%	3.09280	0.02131
wxy	94.20%	1.79961	0.01451	95.00%	3.12948	0.02191
riid	87.60%	1.44443	0.01252	89.40%	2.50063	0.02076
rnid	87.80%	1.44560	0.01254	89.40%	2.50932	0.02085
wiid	94.20%	1.76111	0.01150	93.60%	3.04880	0.01991
wker	97.00%	2.05294	0.00493	97.80%	3.55483	0.00803
wnid	94.40%	1.77719	0.00861	95.40%	3.08834	0.01178

Table C.2: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00160	0.00003
dglm.Tw.0	1.00161	0.00004
glmT0-dglmT0	1.20972	0.00079
robglm.G.log	1.00194	0.00065

Table C.3: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Dispersion Adjusted: Tweedie GLM by dispersion sub-model from Tweedie DGLM, and robust gamma GLM (link=log) conditional variance function estimates.

N_nid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.Gamma.inv_ksm	2.00760	0.02669	4.92326	0.05749
glm.Gamma.inv_N(τ, 1)	2.00708	0.02756	4.92499	0.05769
glm.Tw.0_ksm	2.00812	0.02668	4.92500	0.05744
glm.Tw.0_N(0, 1)	2.00697	0.02746	4.92493	0.05745
dglm.Tw.0_ksm	2.00589	0.02656	4.92647	0.05740
dglm.Tw.0_N(0, 1)	2.00638	0.02745	4.92731	0.05744
glmT0-dglmT0_ksm	2.00548	0.02671	4.92665	0.05744
glmT0-dglmT0_N(0, 1)	2.00501	0.02749	4.92841	0 05743
robglm.G.inv_ksm	2.00725	0.02667	4.92424	0.05752
robglm.G.inv_N(0, 1)	2.00768	0.02745	4.92307	0.05741
robglm.G.log_ksm	2.00670	0.02670	4.92647	0.05758
robglm.G.log_N(0, 1)	2.00528	0.02744	4.93004	0.05736
xy	2.00809	0.02628	4.92861	0.05528
pwy	2.00981	0.02626	4.92454	0.05535
mcmb	2.01265	0.02645	4.91825	0.05567
wxy	2.00795	0.02632	4.92940	0.05540
riid	2.00654	0.02750	4.92598	0.05742
rnid	2.00654	0.02750	4.92598	0.05742
wiid	2.00654	0.02750	4.92598	0.05742
wker	2.00654	0.02750	4.92598	0.05742
wnid	2.00654	0.02750	4.92598	0.05742

Table C.4: Parameter estimates for Model 2 for $\tau = 0.5$.

N_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Gamma.inv_ksm	98.80%	2.84073	0.00731	94.80%	4.92081	0.01259
glm.Gamma.inv_N(0, 1)	96.40%	2.55209	0.00478	96.20%	5.29685	0.00975
glm.Tw.0_ksm	98.60%	2.83792	0.00760	94.80%	4.91312	0.01324
glm.Tw.0_N(0, 1)	95.40%	2.42192	0.00528	96.00%	5.17207	0.00935
dglm.Tw.0_ksm	98.60%	2.83770	0.00761	95.20%	4.91457	0.01307
dglm.Tw.0_N(0, 1)	96.20%	2.49964	0.00585	96.20%	5.19403	0.00905
glmT0-dglmT0_ksm	98.80%	2.84883	0.00755	95.40%	4.93312	0.01334
glmT0-dglmT0_N(0, 1)	91.60%	2.12361	0.00520	91.40%	4.39551	0.00869
robglm.G.inv_ksm	98.80%	2.83991	0.00750	95.20%	4.92115	0.01266
robglm.G.inv_N(0, 1)	96.00%	2.54172	0.00521	95.80%	5.28050	0.01041
robglm.G.log_ksm	98.60%	2.83880	0.00738	95.40%	4.91045	0.01284
robglm.G.log_N(0, 1)	95.20%	2.42341	0.00553	95.80%	5.17850	0.00983
xy	95.40%	2.39756	0.01846	96.60%	5.14327	0.03483
pwy	95.00%	2.40552	0.01816	96.00%	5.16056	0.03475
mcmb	97.40%	2.61812	0.01643	96.40%	5.22662	0.03543
wxy	95.40%	2.39480	0.01836	96.40%	5.14059	0.03491
riid	86.40%	1.89196	0.01607	88.00%	4.00544	0.03290
rnid	86.60%	1.90542	0.01604	89.00%	4.11178	0.03389
wiid	97.60%	2.86566	0.01706	95.60%	4.96099	0.02953
wker	99.00%	2.97660	0.00648	98.40%	5.92896	0.01463
wnid	96.60%	2.41842	0.01281	96.60%	5.16515	0.01873

Table C.5: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Gamma.inv	1.00140	0.00003
glm.Tw.0	1.00142	0.00004
dglm.Tw.0	0.99406	0.00036
glmT0-dglmT0	1.20594	0.00073
robglm.G.inv	1.00417	0.00068
robglm.G.log	1.00224	0.00066

Table C.6: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using gamma GLM (link=inverse), Tweedie GLM (link=log), Tweedie DGLM (link=log), Dispersion Adjusted: Tweedie GLM by dispersion sub-model from Tweedie DGLM, robust gamma GLM (link=inverse) and robust gamma GLM (link=log) conditional variance function estimates.

t(20)_iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	2.00762	0.00504	4.99214	0.00865
glm.Tw.0_t(20)	2.00731	0.00522	4.99201	0.00863
dglm.Tw.0_ksm	2.00750	0.00502	4.99230	0.00863
dglm.Tw.0_t(20)	2.00711	0.00523	4.99254	0.00866
glmT0-dglmT0_ksm	2.00756	0.00504	4.99233	0.00866
glmT0-dglmT0_t(20)	2.00751	0.00523	4.99191	0.00865
robglm.G.log_ksm	2.00756	0.00503	4.99225	0.00864
robglm.G.log_t(20)	2.00694	0.00523	4.99288	0.00865
xy	2.00702	0.00504	4.99421	0.00836
pwy	2.00701	0.00504	4.99399	0.00835
mcmb	2.00651	0.00505	4.99520	0.00838
wxy	2.00688	0.00501	4.99443	0.00833
riid	2.00723	0.00522	4.99234	0.00864
rnid	2.00723	0.00522	4.99234	0.00864
wiid	2.00723	0.00522	4.99234	0.00864
wker	2.00723	0.00522	4.99234	0.00864
wnid	2.00723	0.00522	4.99234	0.00864

Table C.7: Parameter estimates for Model 1 for $\tau = 0.5$.

t(20)_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	94.00%	0.44469	0.00116	95.40%	0.77058	0.00201
glm.Tw.0_t(20)	95.40%	0.46728	0.00085	96.80%	0.80958	0.00142
dglm.Tw.0_ksm	95.00%	0.44439	0.00114	95.80%	0.76922	0.00198
dglm.Tw.0_t(20)	95.40%	0.46885	0.00109	97.00%	0.81016	0.00150
glmT0-dglmT0_ksm	94.40%	0.44544	0.00117	95.60%	0.77139	0.00204
glmT0-dglmT0_t(20)	91.20%	0.38973	0.00092	92.00%	0.67570	0.00128
robglm.G.log_ksm	95.20%	0.44452	0.00113	95.40%	0.77028	0.00192
robglm.G.log_t(20)	94.60%	0.46003	0.00090	96.60%	0.79666	0.00154
xy	94.80%	0.44827	0.00356	96.40%	0.77878	0.00520
pwy	95.60%	0.44971	0.00355	96.60%	0.78064	0.00516
mcmb	94.80%	0.44579	0.00296	95.80%	0.77115	0.00495
wxy	94.80%	0.44789	0.00354	96.80%	0.77755	0.00518
riid	88.80%	0.36160	0.00301	88.40%	0.62902	0.00518
rnid	88.80%	0.36192	0.00301	88.80%	0.63094	0.00516
wiid	93.20%	0.43789	0.00265	94.40%	0.75807	0.00459
wker	98.20%	0.52662	0.00125	98.40%	0.90903	0.00205
wnid	94.40%	0.45320	0.00220	95.80%	0.78184	0.00284

Table C.8: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00159	0.00004
dglm.Tw.0	1.00167	0.00005
glmT0-dglmT0	1.19986	0.00069
robglm.G.log	1.01769	0.00085

Table C.9: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Dispersion Adjusted: Tweedie GLM by dispersion sub-model from Tweedie DGLM, and robust gamma GLM (link=log) conditional variance function estimates.

t(20)_nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	2.01077	0.00665	4.98928	0.01404
glm.Tw.0_t(20)	2.01021	0.00689	4.98966	0.01408
dglm.Tw.0_ksm	2.01086	0.00667	4.98920	0.01407
dglm.Tw.0_t(20)	2.01085	0.00689	4.98856	0.01407
glmT0-dglmT0_ksm	2.01082	0.00667	4.98919	0.01404
glmT0-dglmT0_t(20)	2.01006	0.00688	4.99003	0.01407
xy	2.00903	0.00659	4.99544	0.01358
pwy	2.00909	0.00659	4.99495	0.01355
mcmb	2.00979	0.00661	4.99406	0.01362
wxy	2.00892	0.00656	4.99555	0.01356
riid	2.01038	0.00688	4.98936	0.01405
rnid	2.01038	0.00688	4.98936	0.01405
wiid	2.01038	0.00688	4.98936	0.01405
wker	2.01038	0.00688	4.98936	0.01405
wnid	2.01038	0.00688	4.98936	0.01405

Table C.10: Parameter estimates for Model 2 for $\tau = 0.5$.

t(20)_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	98.60%	0.72071	0.00184	95.40%	1.24750	0.00320
glm.Tw.0_t(20)	96.80%	0.64764	0.00151	97.60%	1.38017	0.00258
dglm.Tw.0_ksm	98.60%	0.72184	0.00191	95.40%	1.25014	0.00328
dglm.Tw.0_t(20)	97.20%	0.66561	0.00172	97.60%	1.38088	0.00257
glmT0-dglmT0_ksm	98.60%	0.72218	0.00189	95.00%	1.25212	0.00320
glmT0-dglmT0_t(20)	92.80%	0.54335	0.00157	93.60%	1.15343	0.00228
xy	93.80%	0.59576	0.00468	95.40%	1.27584	0.00854
pwy	94.60%	0.59732	0.00468	96.00%	1.27792	0.00843
mcmb	97.00%	0.65059	0.00411	96.40%	1.29598	0.00857
wxy	94.80%	0.59511	0.00466	96.00%	1.27417	0.00849
riid	87.00%	0.48710	0.00415	87.80%	1.02230	0.00851
rnid	87.20%	0.49105	0.00412	88.60%	1.04975	0.00856
wiid	97.40%	0.70607	0.00431	94.20%	1.22234	0.00746
wker	98.80%	0.76090	0.00166	99.00%	1.50936	0.00364
wnid	94.40%	0.61118	0.00322	96.60%	1.30388	0.00476

Table C.11: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00142	0.00004
dglm.Tw.0	0.99508	0.00049
glmT0-dglmT0	1.19689	0.00067

Table C.12: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Dispersion Adjusted: Tweedie GLM by dispersion sub-model from Tweedie DGLM, and robust gamma GLM (link=log) conditional variance function estimates.

t(10)_iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.99393	0.00502	5.00590	0.00887
glm.Tw.0_t(10)	1.99250	0.00509	5.00647	0.00888
dglm.Tw.0_ksm	1.99380	0.00503	5.00589	0.00888
dglm.Tw.0_t(10)	1.99282	0.00509	5.00591	0.00886
glmT0-dglmT0_ksm	1.99386	0.00502	5.00609	0.00886
glmT0-dglmT0_t(10)	1.99266	0.00510	5.00619	0.00886
xy	1.99242	0.00497	5.00698	0.00862
pwy	1.99265	0.00497	5.00675	0.00863
mcmb	1.99219	0.00496	5.00749	0.00861
wxy	1.99250	0.00498	5.00692	0.00865
riid	1.99269	0.00509	5.00622	0.00887
rnid	1.99269	0.00509	5.00622	0.00887
wiid	1.99269	0.00509	5.00622	0.00887
wker	1.99269	0.00509	5.00622	0.00887
wnid	1.99269	0.00509	5.00622	0.00887

Table C.13: Parameter estimates for Model 1 for $\tau = 0.5$.

t(10)_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	94.60%	0.45454	0.00119	94.80%	0.78721	0.00207
glm.Tw.0_t(10)	96.60%	0.50094	0.00102	96.40%	0.86813	0.00174
dglm.Tw.0_ksm	94.00%	0.45515	0.00124	95.00%	0.78709	0.00214
dglm.Tw.0_t(10)	96.20%	0.50152	0.00130	96.80%	0.86835	0.00182
glmT0-dglmT0_ksm	94.00%	0.45481	0.00120	94.60%	0.78837	0.00211
glmT0-dglmT0_t(10)	92.80%	0.42211	0.00103	93.00%	0.73065	0.00152
xy	94.40%	0.45742	0.00343	94.40%	0.79652	0.00545
pwy	94.20%	0.46010	0.00345	94.00%	0.80069	0.00542
mcmb	93.80%	0.45503	0.00294	94.00%	0.78871	0.00529
wxy	94.20%	0.45777	0.00339	94.00%	0.79630	0.00532
riid	89.20%	0.37155	0.00315	89.40%	0.64041	0.00517
rnid	89.40%	0.37190	0.00315	89.60%	0.64325	0.00520
wiid	94.40%	0.44750	0.00274	94.40%	0.77470	0.00474
wker	97.40%	0.53896	0.00138	97.80%	0.93363	0.00226
wnid	93.40%	0.45673	0.00228	94.80%	0.79103	0.00286

Table C.14: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00154	0.00003
dglm.Tw.0	1.00162	0.00005
glmT0-dglmT0	1.19079	0.00052

Table C.15: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Dispersion Adjusted: Tweedie GLM by dispersion sub-model from Tweedie DGLM, and robust gamma GLM (link=log) conditional variance function estimates.

t(10)_nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.99488	0.00639	5.01627	0.01378
glm.Tw.0_t(10)	1.99509	0.00658	5.01732	0.01376
dglm.Tw.0_ksm	1.99480	0.00639	5.01666	0.01376
dglm.Tw.0_t(10)	1.99562	0.00657	5.01618	0.01376
glmT0-dglmT0_ksm	1.99479	0.00639	5.01623	0.01376
glmT0-dglmT0_t(10)	1.99564	0.00657	5.01628	0.01374
xy	1.99348	0.00631	5.01965	0.01321
pwy	1.99328	0.00631	5.02028	0.01317
mcmb	1.99311	0.00625	5.02016	0.01321
wxy	1.99278	0.00632	5.02158	0.01320
riid	1.99554	0.00657	5.01647	0.01376
rnid	1.99554	0.00657	5.01647	0.01376
wiid	1.99554	0.00657	5.01647	0.01376
wker	1.99554	0.00657	5.01647	0.01376
wnid	1.99554	0.00657	5.01647	0.01376

Table C.16: Parameter estimates for Model 2 for $\tau = 0.5$.

t(10)_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	98.80%	0.74141	0.00188	96.20%	1.28318	0.00318
glm.Tw.0_t(10)	97.60%	0.69335	0.00168	98.20%	1.48360	0.00297
dglm.Tw.0_ksm	99.00%	0.74191	0.00187	96.20%	1.28494	0.00325
dglm.Tw.0_t(10)	97.80%	0.71444	0.00195	98.00%	1.48726	0.00284
glmT0-dglmT0_ksm	98.80%	0.74373	0.00183	96.20%	1.28854	0.00316
glmT0-dglmT0_t(10)	94.60%	0.58665	0.00179	96.40%	1.25215	0.00282
xy	95.80%	0.62050	0.00494	96.80%	1.33157	0.00902
pwy	96.00%	0.62099	0.00489	97.00%	1.33336	0.00898
mcmb	97.40%	0.67439	0.00431	97.00%	1.34746	0.00924
wxy	95.40%	0.61993	0.00497	96.60%	1.32940	0.00900
riid	90.40%	0.48662	0.00404	91.60%	1.04194	0.00877
rnid	90.80%	0.49035	0.00402	92.40%	1.07156	0.00888
wiid	98.00%	0.72614	0.00423	95.40%	1.25709	0.00732
wker	99.20%	0.78430	0.00176	98.20%	1.56072	0.00369
wnid	95.40%	0.62323	0.00343	96.80%	1.33654	0.00493

Table C.17: Results for Model 2 with $e_i \sim t(10)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00136	0.00003
dglm.Tw.0	0.99479	0.00049
glmT0-dglmT0	1.18932	0.00058

Table C.18: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Dispersion Adjusted: Tweedie GLM by dispersion sub-model from Tweedie DGLM, and robust gamma GLM (link=log) conditional variance function estimates.

Appendix D

Using Non-Parametric Regression Models for Resampling Schemes

N_iid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.97008	0.01912	5.04271	0.03349
glm.Tw.0_N(0, 1)	1.97129	0.01944	5.04104	0.03343
dglm.Tw.0_ksm	1.97029	0.01911	5.04261	0.03348
dglm.Tw.0_N(0, 1)	1.97136	0.01947	5.04128	0.03339
JointMod_ksm	1.97086	0.01909	5.04193	0.03342
JointMod_N(0, 1)	1.97041	0.01948	5.04297	0.03342
LocPoly-p1_ksm	1.97094	0.01909	5.04121	0.03342
LocPoly-p1_N(0, 1)	1.97210	0.01949	5.03913	0.03340
LocPoly-p3_ksm	1.97170	0.01904	5.04025	0.03338
LocPoly-p3_N(0, 1)	1.97084	0.01946	5.04115	0.03346
Lowess_ksm	1.97193	0.01908	5.04004	0.03342
Lowess_N(0, 1)	1.97120	0.01947	5.04163	0.03343
xy	1.97062	0.01872	5.04277	0.03192
pwy	1.97068	0.01879	5.04258	0.03212
mcmb	1.97105	0.01876	5.04194	0.03202
wxy	1.97105	0.01876	5.04219	0.03206
riid	1.97130	0.01946	5.04106	0.03342
rnid	1.97130	0.01946	5.04106	0.03342
wiid	1.97130	0.01946	5.04106	0.03342
wker	1.97130	0.01946	5.04106	0.03342
wnid	1.97130	0.01946	5.04106	0.03342

Table D.1: Parameter estimates for Model 1 for $\tau = 0.5$.

N_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	95.80%	1.75794	0.00478	95.00%	3.04356	0.00825
glm.Tw.0_N(0, 1)	95.80%	1.75096	0.00305	95.40%	3.03294	0.00524
dglm.Tw.0_ksm	96.00%	1.75517	0.00467	95.60%	3.04145	0.00819
dglm.Tw.0_N(0, 1)	96.60%	1.75191	0.00366	95.40%	3.03465	0.00525
JointMod_ksm	96.00%	1.75481	0.00473	95.60%	3.04245	0.00821
JointMod_N(0, 1)	96.40%	1.74938	0.00367	95.20%	3.02716	0.00517
LocPoly-p1_ksm	96.00%	1.75805	0.00465	95.80%	3.04315	0.00821
LocPoly-p1_N(0, 1)	95.80%	1.74641	0.00391	95.20%	3.01999	0.00539
LocPoly-p3_ksm	96.00%	1.75874	0.00472	95.60%	3.04330	0.00826
LocPoly-p3_N(0, 1)	96.40%	1.74132	0.00383	95.00%	3.01230	0.00594
Lowess_ksm	96.00%	1.75496	0.00488	95.40%	3.03819	0.00851
Lowess_N(0, 1)	88.40%	1.33859	0.00375	87.60%	2.31595	0.00607
xy	96.40%	1.80950	0.01406	95.60%	3.14892	0.02135
pwy	95.80%	1.81766	0.01401	95.80%	3.15861	0.02161
mcmb	96.00%	1.80005	0.01201	95.40%	3.11231	0.02130
wxy	96.20%	1.81252	0.01393	95.60%	3.14577	0.02160
riid	90.00%	1.45031	0.01217	90.60%	2.52460	0.02061
rnid	90.20%	1.45135	0.01218	90.60%	2.53046	0.02065
wiid	95.60%	1.76258	0.01106	95.20%	3.05135	0.01915
wker	98.20%	2.06341	0.00473	98.40%	3.56657	0.00740
wnid	96.40%	1.79603	0.00894	96.00%	3.10960	0.01154

Table D.2: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00159	0.00004
dglm.Tw.0	1.00164	0.00004
JointMod	1.00161	0.00004
LocPoly-p1	1.00156	0.00022
LocPoly-p3	1.00105	0.00010
Lowess	1.31318	0.00224

Table D.3: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling of the conditional variance function estimate.

N_iid

Method	mean of the ratio of the determinants	SE
glm.Tw.0_ksm	0.95053	0.00828
glm.Tw.0_N(0, 1)	0.97707	0.01438
dglm.Tw.0_ksm	0.94641	0.00820
dglm.Tw.0_N(0, 1)	0.96930	0.01446
JointMod_ksm	0.94913	0.00838
JointMod_N(0, 1)	0.96454	0.01402
LocPoly-p1_ksm	0.94753	0.00843
LocPoly-p1_N(0, 1)	0.96138	0.01439
LocPoly-p3_ksm	0.94891	0.00816
LocPoly-p3_N(0, 1)	0.94984	0.01397
Lowess_ksm	0.94790	0.00848
Lowess_N(0, 1)	0.33001	0.00505

Table D.4: Results for Model 1 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed and when bootstrapping from the error distribution. Six different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling.

N_nid

Method	β_0		β_1	
	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.95819	0.02538	5.06235	0.05403
glm.Tw.0_N(0, 1)	1.95819	0.02568	5.05920	0.05394
dglm.Tw.0_ksm	1.95824	0.02528	5.06242	0.053980
dglm.Tw.0_N(0, 1)	1.95834	0.02570	5.06006	0.05378
JointMod_ksm	1.96001	0.02528	5.06049	0.05390
JointMod_N(0, 1)	1.95695	0.02572	5.06247	0.05386
LocPoly-p1_ksm	1.95974	0.02529	5.05946	0.05392
LocPoly-p1_N(0, 1)	1.95953	0.02573	5.05614	0.05377
LocPoly-p3_ksm	1.95943	0.02526	5.05906	0.05385
LocPoly-p3_N(0, 1)	1.95770	0.02572	5.05913	0.05396
Lowess_ksm	1.96154	0.02516	5.05761	0.05391
Lowess_N(0, 1)	1.95813	0.02572	5.06035	0.05392
xy	1.96024	0.02476	5.06048	0.05213
pwy	1.96023	0.02489	5.06029	0.05246
mcmb	1.95822	0.02475	5.06179	0.05240
wxy	1.96062	0.02481	5.05982	0.05236
riid	1.95824	0.02570	5.05940	0.05388
rnid	1.95824	0.02570	5.05940	0.05388
wiid	1.95824	0.02570	5.05940	0.05388
wker	1.95824	0.02570	5.05940	0.05388
wnid	1.95824	0.02570	5.05940	0.05388

Table D.5: Parameter estimates for Model 2 for $\tau = 0.5$.

N_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	98.80%	2.84800	0.00782	95.20%	4.93183	0.01349
glm.Tw.0_N(0, 1)	96.20%	2.41679	0.00518	96.60%	5.17082	0.00897
dglm.Tw.0_ksm	98.60%	2.84679	0.00761	95.20%	4.93368	0.01337
dglm.Tw.0_N(0, 1)	96.60%	2.50084	0.00586	97.40%	5.19316	0.00946
JointMod_ksm	98.60%	2.84149	0.00787	95.20%	4.92621	0.01362
JointMod_N(0, 1)	95.60%	2.31986	0.00552	95.80%	4.96387	0.00852
LocPoly-p1_ksm	98.60%	2.84874	0.00768	95.20%	4.93199	0.01354
LocPoly-p1_N(0, 1)	95.00%	2.31029	0.00620	95.20%	4.94583	0.00895
LocPoly-p3_ksm	98.80%	2.99623	0.14745	95.20%	5.17381	0.24556
LocPoly-p3_N(0, 1)	95.80%	2.30193	0.00628	95.40%	4.93398	0.01035
Lowess_ksm	98.60%	2.85382	0.00789	96.00%	4.94030	0.01375
Lowess_N(0, 1)	89.80%	1.91826	0.00500	86.80%	3.66798	0.00932
xy	96.20%	2.40816	0.01858	96.40%	5.17043	0.03610
pwy	96.20%	2.42093	0.01850	96.20%	5.18461	0.03662
mcmb	96.60%	2.63146	0.01680	96.00%	5.24005	0.03734
wxy	96.20%	2.41214	0.01842	95.80%	5.16198	0.03678
riid	90.20%	1.89597	0.01626	88.80%	4.01990	0.03319
rnid	90.20%	1.90844	0.01616	89.80%	4.11900	0.03380
wiid	97.40%	2.85299	0.01787	95.80%	4.93905	0.03094
wker	99.20%	2.98197	0.00628	98.60%	5.94463	0.01376
wnid	96.40%	2.43033	0.01324	96.20%	5.19870	0.01933

Table D.6: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00144	0.00003
dglm.Tw.0	0.99349	0.00034
JointMod	1.00206	0.00024
LocPoly-p1	1.00263	0.00027
LocPoly-p3	1.24693	0.24497
Lowess	1.36807	0.00237

Table D.7: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling of the conditional variance function estimate.

N_nid

Method	mean of the ratio of the determinants	SE
glm.Tw.0_ksm	0.93684	0.00875
glm.Tw.0_N(0, 1)	1.02362	0.01594
dglm.Tw.0_ksm	0.93755	0.00879
dglm.Tw.0_N(0, 1)	1.07035	0.01701
JointMod_ksm	0.93355	0.00890
JointMod_N(0, 1)	0.92790	0.01428
LocPoly-p1_ksm	0.93364	0.00878
LocPoly-p1_N(0, 1)	0.92035	0.01440
LocPoly-p3_ksm	1521.29203	1520.35943
LocPoly-p3_N(0, 1)	0.90958	0.01431
Lowess_ksm	0.94913	0.00960
Lowess_N(0, 1)	0.29859	0.00467

Table D.8: Results for Model 2 with $e_i \sim \mathcal{N}(0, 16)$ for $\tau = 0.5$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed and when bootstrapping from the error distribution. Six different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modellnig, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling.

t(20)_iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.99891	0.00505	5.00030	0.00889
glm.Tw.0_t(20)	1.99901	0.00515	5.00076	0.00888
dglm.Tw.0_ksm	1.99854	0.00504	5.00081	0.00887
dglm.Tw.0_t(20)	1.99947	0.00514	5.00041	0.00889
JointMod_ksm	1.99882	0.00505	5.00061	0.00889
JointMod_t(20)	1.99896	0.00514	5.00093	0.00888
LocPoly-p1_ksm	1.99875	0.00506	5.00045	0.008890
LocPoly-p1_t(20)	1.99910	0.00515	5.00093	0.00890
LocPoly-p3_ksm	1.99856	0.00506	5.00075	0.00890
LocPoly-p3_t(20)	1.99941	0.00514	5.00040	0.00888
Lowess_ksm	1.99871	0.00506	5.00071	0.00889
Lowess_t(20)	1.99937	0.00515	5.00048	0.00888
xy	1.99968	0.00497	4.99859	0.00858
pwy	1.99943	0.00497	4.99878	0.00856
mcmb	1.99932	0.00498	4.99881	0.00861
wxy	1.99959	0.00497	4.99890	0.00857
riid	1.99931	0.00515	5.00059	0.00889
rnid	1.99931	0.00515	5.00059	0.00889
wiid	1.99931	0.00515	5.00059	0.00889
wker	1.99931	0.00515	5.00059	0.00889
wnid	1.99931	0.00515	5.00059	0.00889

Table D.9: Parameter estimates for Model 1 for $\tau = 0.5$.

t(20)_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	94.00%	0.44431	0.00116	95.00%	0.76895	0.00201
glm.Tw.0_t(20)	95.40%	0.46573	0.00086	95.80%	0.80652	0.00148
dglm.Tw.0_ksm	94.80%	0.44340	0.00114	94.60%	0.76824	0.00195
dglm.Tw.0_t(20)	95.20%	0.46616	0.00104	95.80%	0.80702	0.00143
JointMod_ksm	94.80%	0.44343	0.00113	95.40%	0.76830	0.00193
JointMod_t(20)	95.20%	0.46441	0.00105	95.40%	0.80478	0.00148
LocPoly-p1_ksm	94.40%	0.44342	0.00115	95.40%	0.76780	0.00198
LocPoly-p1_t(20)	95.00%	0.46463	0.00112	96.20%	0.80539	0.00170
LocPoly-p3_ksm	94.80%	0.44369	0.00114	95.40%	0.76794	0.00195
LocPoly-p3_t(20)	94.60%	0.46252	0.00110	95.60%	0.80172	0.00181
Lowess_ksm	93.80%	0.44312	0.00113	95.00%	0.76833	0.00199
Lowess_t(20)	85.20%	0.34224	0.00098	85.20%	0.59254	0.00164
xy	93.60%	0.44949	0.00338	95.80%	0.78612	0.00507
pwy	94.00%	0.45060	0.00340	95.40%	0.78711	0.00513
mcmb	94.00%	0.44745	0.00287	95.60%	0.77752	0.00494
wxy	93.60%	0.44913	0.00342	96.00%	0.78514	0.00510
riid	89.40%	0.36782	0.00299	90.20%	0.63665	0.00512
rnid	89.40%	0.36810	0.00300	90.40%	0.63856	0.00511
wiid	92.80%	0.44010	0.00261	94.40%	0.76189	0.00451
wker	97.60%	0.52440	0.00123	97.60%	0.90818	0.00209
wnid	97.40%	0.44924	0.00218	95.40%	0.77975	0.00284

Table D.10: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00156	0.00003
dglm.Tw.0	1.00161	0.00004
JointMod	1.00159	0.00005
LocPoly-p1	1.00100	0.00026
LocPoly-p3	1.00114	0.00010
Lowess	1.36707	0.00263

Table D.11: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling of the conditional variance function estimate.

t(20)_iid

Method	mean of the ratio of the determinants	SE
glm.Tw.0_ksm	0.98682	0.00927
glm.Tw.0_t(20)	1.23955	0.01869
dglm.Tw.0_ksm	0.97809	0.00902
dglm.Tw.0_t(20)	1.23406	0.01868
JointMod_ksm	0.98199	0.00912
JointMod_t(20))	1.22392	0.01840
LocPoly-p1_ksm	0.97818	0.00925
LocPoly-p1_t(20)	1.22365	0.01859
LocPoly-p3_ksm	0.97826	0.00909
LocPoly-p3_t(20)	1.20631	0.01843
Lowess_ksm	0.97652	0.00887
Lowess_t(20)	0.35914	0.00547

Table D.12: Results for Model 1 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed and when bootstrapping from the error distribution. Six different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling.

t(20)_nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.99835	0.00667	5.00041	0.01420
glm.Tw.0_t(20)	1.99859	0.00678	5.00110	0.01418
dglm.Tw.0_ksm	1.99772	0.00666	5.00130	0.01418
dglm.Tw.0_t(20)	1.99920	0.00677	5.00072	0.01422
JointMod_ksm	1.99814	0.00666	5.00095	0.01421
JointMod_t(20)	1.99853	0.00678	5.00130	0.01420
LocPoly-p1_ksm	1.99800	0.00667	5.00071	0.01419
LocPoly-p1_t(20)	1.99869	0.00679	5.00152	0.01422
LocPoly-p3_ksm	1.99744	0.00670	5.00125	0.01422
LocPoly-p3_t(20)	1.99909	0.00678	5.00075	0.01420
Lowess_ksm	1.99781	0.00665	5.00119	0.01420
Lowess_t(20)	1.99903	0.00678	5.00085	0.01418
xy	1.99984	0.00658	4.99662	0.01387
pwy	1.99955	0.00658	4.99672	0.01385
mcmb	1.99957	0.00659	4.99777	0.01386
wxy	1.99969	0.00658	4.99716	0.01386
riid	1.99895	0.00678	5.00101	0.01419
rnid	1.99895	0.00678	5.00101	0.01419
wiid	1.99895	0.00678	5.00101	0.01419
wker	1.99895	0.00678	5.00101	0.01419
wnid	1.99895	0.00678	5.00101	0.01419

Table D.13: Parameter estimates for Model 2 for $\tau = 0.5$.

t(20)_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	98.40%	0.71913	0.00190	95.00%	1.24459	0.00328
glm.Tw.0_t(20)	96.20%	0.64248	0.00150	97.80%	1.37575	0.00260
dglm.Tw.0_ksm	98.40%	0.71855	0.00182	95.40%	1.24535	0.00315
dglm.Tw.0_t(20)	97.00%	0.66393	0.00176	97.60%	1.37970	0.00257
JointMod_ksm	98.60%	0.71778	0.00186	95.60%	1.24312	0.00319
JointMod_t(20)	95.20%	0.61618	0.00160	96.80%	1.32020	0.00250
LocPoly-p1_ksm	98.60%	0.71757	0.00186	95.00%	1.24227	0.00321
LocPoly-p1_t(20)	94.80%	0.61291	0.00182	96.80%	1.31649	0.00282
LocPoly-p3_ksm	98.00%	0.73165	0.01016	95.40%	1.26658	0.01739
LocPoly-p3_t(20)	95.80%	0.61117	0.00176	96.20%	1.31399	0.00318
Lowess_ksm	98.80%	0.72042	0.00180	95.20%	1.24904	0.00316
Lowess_t(20)	88.00%	0.49395	0.00129	83.40%	0.94450	0.00250
xy	93.60%	0.59822	0.00450	96.00%	1.29107	0.00849
pwy	94.20%	0.59963	0.00451	95.40%	1.29236	0.00857
mcmb	95.80%	0.65479	0.00404	96.40%	1.31082	0.00869
wxy	93.60%	0.59793	0.00454	95.80%	1.28999	0.00847
riid	90.20%	0.48638	0.00422	89.00%	1.01986	0.00830
rnid	91.00%	0.48980	0.00420	90.40%	1.04729	0.00851
wiid	98.20%	0.71198	0.00423	95.00%	1.23257	0.00732
wker	98.80%	0.75884	0.00163	98.40%	1.51193	0.00383
wnid	95.00%	0.60564	0.00324	96.00%	1.30278	0.00480

Table D.14: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00139	0.00004
dglm.Tw.0	0.99477	0.00050
JointMod	1.00245	0.00029
LocPoly-p1	1.00278	0.00032
LocPoly-p3	1.07827	0.05557
Lowess	1.41872	0.00264

Table D.15: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling of the conditional variance function estimate.

t(20)_nid

Method	mean of the ratio of the determinants	SE
glm.Tw.0_ksm	0.97580	0.00940
glm.Tw.0_t(10)	1.30935	0.02045
dglm.Tw.0_ksm	0.97286	0.00926
dglm.Tw.0_N(τ, 1)	1.36783	0.02184
JointMod_ksm	0.96918	0.00922
JointMod_t(20))	1.18451	0.01830
LocPoly-p1_ksm	0.96617	0.00943
LocPoly-p1_t(10)	1.17147	0.01857
LocPoly-p3_ksm	8.33778	6.85549
LocPoly-p3_t(10)	1.16455	0.01883
Lowess_ksm	0.98504	0.00960
Lowess_t(10)	0.33394	0.00508

Table D.16: Results for Model 2 with $e_i \sim t(20)$ for $\tau = 0.5$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed and when bootstrapping from the error distribution. Six different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie GLM (link=log), Tweedie DGLM (link=log), Joint Modelling, Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling.

t(10)_iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.99635	0.00454	5.00649	0.00830
glm.Tw.0_t(10)	1.99458	0.00472	5.00647	0.00836
dglm.Tw.0_ksm	1.99617	0.00455	5.00675	0.00832
dglm.Tw.0_t(10)	1.99480	0.00471	5.00602	0.00834
LocPoly-p1_ksm	1.99608	0.00453	5.00686	0.00832
LocPoly-p1_t(10)	1.99459	0.00472	5.00625	0.00833
LocPoly-p3_ksm	1.99619	0.00455	5.00659	0.00833
LocPoly-p3_t(10)	1.99484	0.00470	5.00573	0.00833
Lowess_ksm	1.99659	0.00454	5.00616	0.00834
Lowess_t(10)	1.99447	0.00471	5.00668	0.00830
xy	1.99586	0.00453	5.00445	0.00798
pwy	1.99564	0.00454	5.00471	0.00799
mcmb	1.99486	0.00453	5.00608	0.00799
wxy	1.99598	0.00453	5.00446	0.00799
riid	1.99458	0.00471	5.00646	0.00832
rnid	1.99458	0.00471	5.00646	0.00832
wiid	1.99458	0.00471	5.00646	0.00832
wker	1.99458	0.00471	5.00646	0.00832
wnid	1.99458	0.00471	5.00646	0.00832

Table D.17: Parameter estimates for Model 1 for $\tau = 0.5$.

t(10)_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	97.00%	0.45449	0.00117	94.80%	0.78730	0.00203
glm.Tw.0_t(10)	97.60%	0.50224	0.00101	97.00%	0.86923	0.00175
dglm.Tw.0_ksm	96.80%	0.45422	0.00117	95.40%	0.78623	0.00201
dglm.Tw.0_t(10)	97.80%	0.50055	0.00126	97.60%	0.86751	0.00170
LocPoly-p1_ksm	96.80%	0.45334	0.00119	95.80%	0.78520	0.00203
LocPoly-p1_t(10)	97.80%	0.49887	0.00132	97.40%	0.86526	0.00198
LocPoly-p3_ksm	97.00%	0.45288	0.00116	95.60%	0.78422	0.00199
LocPoly-p3_t(10)	97.20%	0.49591	0.00137	97.00%	0.85874	0.00223
Lowess_ksm	97.20%	0.45398	0.00119	95.40%	0.78619	0.00209
Lowess_t(10)	88.80%	0.35365	0.00099	90.00%	0.61203	0.00162
xy	95.60%	0.45886	0.00370	96.80%	0.80256	0.00569
pwy	95.80%	0.46040	0.00367	96.40%	0.80460	0.00568
mcmb	95.80%	0.45671	0.00312	96.60%	0.79363	0.00551
wxy	95.80%	0.45836	0.00367	96.60%	0.79942	0.00566
riid	89.00%	0.36665	0.00294	88.60%	0.63751	0.00495
rnid	89.20%	0.36699	0.00294	89.00%	0.63964	0.00494
wiid	95.60%	0.44871	0.00285	94.60%	0.77681	0.00494
wker	98.40%	0.53831	0.00135	98.40%	0.93195	0.00214
wnid	95.80%	0.45478	0.00229	95.40%	0.78850	0.00299

Table D.18: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00151	0.00003
dglm.Tw.0	1.00159	0.00005
LocPoly-p1	1.00134	0.00028
LocPoly-p3	1.00106	0.00019
Lowess	1.42122	0.00276

Table D.19: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling of the conditional variance function estimate.

t(10)_iid

Method	mean of the ratio of the determinants	SE
glm.Tw.0_ksm	1.03840	0.01034
glm.Tw.0_t(10)	1.60469	0.02665
dglm.Tw.0_ksm	1.03254	0.01032
dglm.Tw.0_N(τ, 1)	1.58352	0.02606
LocPoly-p1_ksm	1.02804	0.01002
LocPoly-p1_t(10)	1.57245	0.02676
LocPoly-p3_ksm	1.02557	0.00992
LocPoly-p3_t(10)	1.53534	0.02607
Lowess_ksm	1.02983	0.01008
Lowess_t(10)	0.39162	0.00598

Table D.20: Results for Model 1 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed and when bootstrapping from the error distribution. Five different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie GLM (link=log), Tweedie DGLM (link=log), Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling.

t(10)_nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
glm.Tw.0_ksm	1.99622	0.00607	5.00682	0.01396
glm.Tw.0_t(10)	1.99279	0.00632	5.00681	0.01406
dglm.Tw.0_ksm	1.99594	0.00608	5.00718	0.01399
dglm.Tw.0_t(10)	1.99318	0.00630	5.00599	0.01403
LocPoly-p1_ksm	1.99597	0.00606	5.00731	0.01400
LocPoly-p1_t(10)	1.99292	0.00630	5.00626	0.01400
LocPoly-p3_ksm	1.99550	0.00605	5.00680	0.01400
LocPoly-p3_t(10)	1.99325	0.00628	5.00546	0.01402
Lowess_ksm	1.99678	0.00606	5.00626	0.01402
Lowess_t(10)	1.99269	0.00628	5.00706	0.01396
xy	1.99487	0.00597	5.00461	0.01346
pwy	1.99466	0.00598	5.00488	0.01349
mcmb	1.99407	0.00604	5.00664	0.01366
wxy	1.99505	0.00598	5.00470	0.01351
riid	1.99285	0.00629	5.00673	0.01399
rnid	1.99285	0.00629	5.00673	0.01399
wiid	1.99285	0.00629	5.00673	0.01399
wker	1.99285	0.00629	5.00673	0.01399
wnid	1.99285	0.00629	5.00673	0.01399

Table D.21: Parameter estimates for Model 2 for $\tau = 0.5$.

t(10)_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
glm.Tw.0_ksm	99.00%	0.73614	0.00192	95.80%	1.27500	0.00334
glm.Tw.0_t(10)	98.00%	0.69239	0.00174	98.20%	1.48152	0.00304
dglm.Tw.0_ksm	98.60%	0.73701	0.00192	96.20%	1.27552	0.00330
dglm.Tw.0_t(10)	98.40%	0.71355	0.00206	98.20%	1.48365	0.00301
LocPoly-p1_ksm	98.80%	0.73371	0.00198	96.00%	1.27120	0.00338
LocPoly-p1_t(10)	96.60%	0.65834	0.00219	97.60%	1.41597	0.00327
LocPoly-p3_ksm	99.00%	0.77743	0.02292	96.80%	1.34546	0.03939
LocPoly-p3_t(10)	96.80%	0.65147	0.00234	97.40%	1.40234	0.00395
Lowess_ksm	99.00%	0.73909	0.00192	95.40%	1.28026	0.00334
Lowess_t(10)	91.40%	0.51018	0.00134	87.40%	0.97460	0.00247
xy	95.80%	0.61134	0.00494	96.20%	1.31804	0.00941
pwy	95.60%	0.61341	0.00490	96.00%	1.32070	0.00942
mcmb	96.80%	0.66862	0.00443	96.40%	1.33645	0.00963
wxy	95.60%	0.61040	0.00489	96.20%	1.31221	0.00933
riid	90.00%	0.48841	0.00419	89.80%	1.02951	0.00802
rnid	90.20%	0.49196	0.00416	90.80%	1.05965	0.00838
wiid	98.00%	0.72594	0.00463	94.80%	1.25674	0.00802
wker	99.20%	0.77821	0.00179	98.40%	1.54669	0.00376
wnid	96.00%	0.61407	0.00342	96.00%	1.31932	0.00506

Table D.22: Results for Model 2 with $e_i \sim t(10)$ for $\tau = 0.5$. Column C is coverage probability and column L is average length of the 95% confidence intervals for each coefficient.

Method	mean variance of the standardised residuals	SE
glm.Tw.0	1.00138	0.00003
dglm.Tw.0	0.99495	0.00063
LocPoly-p1	1.00340	0.00041
LocPoly-p3	1.23384	0.12002
Lowess	1.47436	0.00288

Table D.23: Results for Model 2 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean variance of the standardised residuals when using Tweedie GLM (link=log), Tweedie DGLM (link=log), Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling of the conditional variance function estimate.

t(10)_nid

Method	mean of the ratio of the determinants	SE
glm.Tw.0_ksm	1.02109	0.01041
glm.Tw.0_t(10)	1.67756	0.02845
dglm.Tw.0_ksm	1.02266	0.01049
dglm.Tw.0_t(10)	1.73832	0.02951
LocPoly-p1_ksm	1.00866	0.01002
LocPoly-p1_t(10)	1.49590	0.02597
LocPoly-p3_ksm	69.63423	53.67201
LocPoly-p3_t(10)	1.45510	0.02614
Lowess_ksm	1.03563	0.01068
Lowess_t(10)	0.36277	0.00574

Table D.24: Results for Model 2 with $e_i \sim t(10)$ for $\tau = 0.5$. Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed and when bootstrapping from the error distribution. Five different methods are used for modeling the conditional variance function for the resampling scheme: Tweedie GLM (link=log), Tweedie DGLM (link=log), Local Polynomial of $p = 1$, Local Polynomial of $p = 3$ and Lowess modeling.

Appendix E

Difference-Based Variance Function Estimation

Homoscedastic, Location Shift Model

The results of the problem of estimating the 50^{th} ($\tau = 0.5$) quantile function of the homoscedastic, location-shift model (M1):

$$\text{Model 1} \quad : \quad y_i = 2 + 5x_i + e_i \quad \text{and}$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$ are presented i the following tables. The error, $\{e_i\}$, is iid from four different distributions: $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$.

$e \sim \mathcal{N}(0, 0.04)$ _iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00076	0.00100	4.99814	0.00176
LocPoly-p1-res	2.00077	0.00100	4.99816	0.00176
LocPoly-p1-diff	2.00080	0.00100	4.99812	0.00176
LocPoly-p3-diff	2.00070	0.00099	4.99829	0.00176
xy	2.00123	0.00098	4.99780	0.00169
pwy	2.00113	0.00098	4.99794	0.00169
mcmb	2.00112	0.00099	4.99794	0.00170
wxy	2.00121	0.00098	4.99775	0.00169
riid	2.00101	0.00102	4.99819	0.00176
rnid	2.00101	0.00102	4.99819	0.00176
wiid	2.00101	0.00102	4.99819	0.00176
wker	2.00101	0.00102	4.99819	0.00176
wnid	2.00101	0.00102	4.99819	0.00176

Table E.1: Parameter estimates for Model 1 for $\tau = 0.5$.

$e \sim \mathcal{N}(0, 0.04)$_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	96.00%	0.08735	0.00023	94.00%	0.15127	0.00040
LocPoly-p1-res	96.00%	0.08732	0.00022	94.00%	0.15112	0.00040
LocPoly-p1-diff	95.60%	0.08721	0.00023	94.20%	0.15109	0.00040
LocPoly-p3-diff	95.80%	0.08743	0.00024	94.40%	0.15131	0.00041
xy	95.80%	0.09010	0.00071	95.40%	0.15704	0.00109
pwy	95.40%	0.09032	0.00071	95.60%	0.15752	0.00109
mcmb	95.60%	0.08946	0.00060	94.00%	0.15539	0.00105
wxy	95.40%	0.08986	0.00070	95.40%	0.15684	0.00108
riid	88.60%	0.07169	0.00405	88.40%	0.12481	0.00703
rnid	88.60%	0.07174	0.00405	88.60%	0.12532	0.00703
wiid	92.60%	0.08746	0.00054	93.80%	0.15140	0.00094
wker	98.20%	0.10289	0.00023	98.00%	0.17772	0.00037
wnid	94.60%	0.08866	0.00043	94.80%	0.15404	0.00057

Table E.2: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim \mathcal{N}(0, 16)$_iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.01516	0.019934	4.96271	0.03520
LocPoly-p1-res	2.01534	0.01996	4.96314	0.03527
LocPoly-p1-diff	2.01603	0.01991	4.96234	0.03515
LocPoly-p3-diff	2.01623	0.02002	4.96422	0.03518
xy	2.02466	0.01965	4.95596	0.03383
pwy	2.02264	0.01961	4.95868	0.03377
mcmb	2.02356	0.01974	4.95615	0.03405
wxy	2.02414	0.01960	4.95502	0.03371
riid	2.02013	0.02032	4.96377	0.03518
rnid	2.02013	0.02032	4.96377	0.03518
wiid	2.02013	0.02032	4.96377	0.03518
wker	2.02013	0.02032	4.96377	0.03518
wnid	2.02013	0.02032	4.96377	0.03518

Table E.3: Parameter estimates for Model 1 for $\tau = 0.5$.

$e \sim \mathcal{N}(0, 16)$_iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	95.80%	1.74695	0.00462	94.00%	3.02543	0.00796
LocPoly-p1-res	96.00%	1.74646	0.00450	94.00%	3.02244	0.00777
LocPoly-p1-diff	95.60%	1.74419	0.00451	94.20%	3.02186	0.00790
LocPoly-p3-diff	95.80%	1.83965	0.06067	94.40%	3.18403	0.10493
xy	95.80%	1.80204	0.01416	95.40%	3.14087	0.02173
pwy	95.40%	1.80643	0.01412	95.60%	3.15037	0.02178
mcmb	95.40%	1.78975	0.01202	94.20%	3.10668	0.02107
wxy	95.40%	1.79668	0.01407	95.40%	3.13632	0.02151
riid	87.80%	1.44394	0.01201	88.60%	2.50939	0.01970
rnid	88.00%	1.44504	0.01202	88.60%	2.51729	0.01974
wiid	92.60%	1.74910	0.01088	93.80%	3.02802	0.01883
wker	98.20%	2.05776	0.00464	98.00%	3.55450	0.00737
wnid	94.60%	1.77310	0.00852	94.80%	3.08072	0.01146

Table E.4: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim t(20)_$iid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00444	0.00504	4.99074	0.00900
LocPoly-p1-res	2.00421	0.00502	4.99129	0.00896
LocPoly-p1-diff	2.00424	0.00502	4.99125	0.00894
LocPoly-p3-diff	2.00437	0.00505	4.99102	0.00897
xy	2.00395	0.00500	4.99117	0.00862
pwy	2.00405	0.00499	4.99080	0.00862
mcmb	2.00395	0.00502	4.99090	0.00868
wxy	2.00432	0.00500	4.99080	0.00864
riid	2.00325	0.00521	4.99136	0.00896
rnid	2.00325	0.00521	4.99136	0.00896
wiid	2.00325	0.00521	4.99136	0.00896
wker	2.00325	0.00521	4.99136	0.00896
wnid	2.00325	0.00521	4.99136	0.00896

Table E.5: Parameter estimates for Model 1 for $\tau = 0.5$.

$e \sim t(20)_$iid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	94.80%	0.44459	0.00121	95.40%	0.77045	0.00212
LocPoly-p1-res	95.40%	0.44476	0.00119	95.20%	0.77045	0.00211
LocPoly-p1-diff	95.20%	0.44552	0.00120	95.60%	0.77084	0.00206
LocPoly-p3-diff	95.00%	0.45387	0.00668	95.00%	0.78582	0.01140
xy	94.00%	0.45181	0.00347	94.80%	0.78553	0.00538
pwy	94.20%	0.45249	0.00350	95.40%	0.78567	0.00539
mcmb	93.80%	0.44914	0.00302	94.80%	0.77732	0.00525
wxy	94.00%	0.45181	0.00348	95.00%	0.78509	0.00542
riid	89.00%	0.36388	0.00301	89.60%	0.63370	0.00523
rnid	89.00%	0.36419	0.00302	89.60%	0.63571	0.00523
wiid	94.00%	0.44151	0.00277	93.40%	0.76433	0.00479
wker	98.00%	0.52624	0.00127	98.00%	0.91058	0.00208
wnid	95.00%	0.45211	0.00228	95.00%	0.78047	0.00297

Table E.6: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim t(10)\_\text{iid}$

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00401	0.00502	4.99995	0.00879
LocPoly-p1-res	2.00410	0.00499	5.00002	0.00876
LocPoly-p1-diff	2.00400	0.00500	5.00000	0.00878
LocPoly-p3-diff	2.00364	0.00498	5.00091	0.00875
xy	2.00422	0.00487	4.99967	0.00836
pwy	2.00426	0.00487	4.99948	0.00836
mcmb	2.00385	0.00488	5.00018	0.00840
wxy	2.00445	0.00488	4.99918	0.00838
riid	2.00414	0.00506	5.00036	0.00877
rnid	2.00414	0.00506	5.00036	0.00877
wiid	2.00414	0.00506	5.00036	0.00877
wker	2.00414	0.00506	5.00036	0.00877
wnid	2.00414	0.00506	5.00036	0.00877

Table E.7: Parameter estimates for Model 1 for $\tau = 0.5$.

$e \sim t(10)\_\text{iid}$

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	96.60%	0.45593	0.00115	95.80%	0.78852	0.00201
LocPoly-p1-res	97.40%	0.45574	0.00117	96.20%	0.78934	0.00202
LocPoly-p1-diff	97.00%	0.45560	0.00116	96.60%	0.78983	0.00201
LocPoly-p3-diff	97.20%	0.46357	0.00489	95.80%	0.80230	0.00848
xy	96.80%	0.46707	0.00373	96.60%	0.81146	0.00595
pwy	96.80%	0.46865	0.00375	96.40%	0.81392	0.00595
mcmb	96.80%	0.46302	0.00317	95.60%	0.80118	0.00576
wxy	96.80%	0.46650	0.00375	96.00%	0.81040	0.00599
riid	91.40%	0.37286	0.00311	91.8%	0.64929	0.00565
rnid	91.40%	0.37320	0.00311	91.8%	0.65126	0.00564
wiid	95.00%	0.45247	0.00284	95.80%	0.78330	0.00492
wker	99.60%	0.54050	0.00134	99.20%	0.93600	0.00221
wnid	96.00%	0.45815	0.00217	96.40%	0.79334	0.00285

Table E.8: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

Method	mean variance of the standardised residuals	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	1.00166	0.00008
LocPoly-p1-ress	1.00180	0.00024
LocPoly-p1-diff	1.00298	0.00103
LocPoly-p3-diff	1.01090	0.00451
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	1.00166	0.00008
LocPoly-p1-ress	1.00180	0.00024
LocPoly-p1-diff	1.00298	0.00103
LocPoly-p3-diff	1.20746	0.13027
$e \sim t(20)$		
dglm.Tw.0	1.00186	0.00010
LocPoly-p1-ress	1.00152	0.00024
LocPoly-p1-diff	1.00125	0.00102
LocPoly-p3-diff	1.08406	0.05329
$e \sim t(10)$		
dglm.Tw.0	1.00182	0.00011
LocPoly-p1-ress	1.00070	0.00028
LocPoly-p1-diff	0.99923	0.00104
LocPoly-p3-diff	1.08381	0.04137

Table E.9: Average values of the variance of the standardised residuals when applying the conditional variance estimate obtained by Tweedie DGLM (link=log) model, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach and Local Polynomial of $p = 3$ difference-based approach.

Method	mean of the ratio of the determinants	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	0.96631	0.00854
LocPoly-p1-ress	0.96131	0.00855
LocPoly-p1-diff	0.96296	0.00842
LocPoly-p3-diff	0.96414	0.00873
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	0.96633	0.00854
LocPoly-p1-ress	0.96129	0.00855
LocPoly-p1-diff	0.96298	0.00842
LocPoly-p3-diff	115.65454	108.04539
$e \sim t(20)$		
dglm.Tw.0	0.98367	0.00836
LocPoly-p1-ress	0.98316	0.00813
LocPoly-p1-diff	0.98732	0.00837
LocPoly-p3-diff	6.88068	4.96824
$e \sim t(10)$		
dglm.Tw.0	1.01626	0.00965
LocPoly-p1-ress	1.01890	0.00975
LocPoly-p1-diff	1.01521	0.00989
LocPoly-p3-diff	2.07418	0.58063

Table E.10: Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed. The conditional variance is estimated by using Tweedie DGLM (link=log) model, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach and Local Polynomial of $p = 3$ difference-based approach.

Heteroscedastic, Location-Scale Shift Model With Linear Variance Function

The results of the problem of estimating the 50^{th} ($\tau = 0.5$) quantile function of the heteroscedastic, location-scale shift model (M2):

$$\text{Model 2} \quad : \quad y_i = 2 + 5x_i + \sigma(x_i)e_i$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$ are presented i the following tables. The variance function is given by

$$\sigma^2(x_i) = 1 + 4x_i,$$

and the error, $\{e_i\}$, is iid from four different distributions: $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$.

$e \sim \mathcal{N}(0, 0.04)$ _nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00119	0.00134	4.99664	0.00287
LocPoly-p1-res	2.00119	0.00134	4.99672	0.00288
LocPoly-p1-diff	2.00126	0.00134	4.99662	0.00287
LocPoly-p3-diff	2.00112	0.00134	4.99687	0.00287
xy	2.00172	0.00131	4.99661	0.00276
pwy	2.00159	0.00130	4.99679	0.00276
mcmb	2.00163	0.00131	4.99677	0.00276
wxy	2.00171	0.00130	4.99644	0.00275
riid	2.00165	0.00135	4.99672	0.00287
rnid	2.00165	0.00135	4.99672	0.00287
wiid	2.00165	0.00135	4.99672	0.00287
wker	2.00165	0.00135	4.99672	0.00287
wnid	2.00165	0.00135	4.99672	0.00287

Table E.11: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim \mathcal{N}(0, 0.04)$_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.60%	0.14161	0.00038	94.00%	0.24520	0.00065
LocPoly-p1-res	98.60%	0.14141	0.00037	93.60%	0.24475	0.00064
LocPoly-p1-diff	98.60%	0.14115	0.00037	93.80%	0.24466	0.00065
LocPoly-p3-diff	98.80%	0.14306	0.00074	94.00%	0.24763	0.00128
xy	94.80%	0.11990	0.00093	95.00%	0.25717	0.00179
pwy	95.40%	0.12017	0.00093	95.60%	0.25802	0.00180
mcmb	97.00%	0.13082	0.00084	96.40%	0.26147	0.00183
wxy	95.60%	0.11956	0.00092	96.00%	0.25706	0.00178
riid	89.00%	0.09520	0.00082	85.20%	0.20163	0.00167
rnid	89.20%	0.09584	0.00081	86.40%	0.20703	0.00170
wiid	97.80%	0.14166	0.00088	92.40%	0.24524	0.00152
wker	99.20%	0.14875	0.00030	98.40%	0.29621	0.00067
wnid	94.60%	0.11953	0.00062	95.20%	0.25753	0.00096

Table E.12: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim \mathcal{N}(0, 16)$_nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	1.97772	0.02646	5.05603	0.05890
LocPoly-p1-res	1.97694	0.02656	5.05830	0.05904
LocPoly-p1-diff	1.97730	0.02653	5.05641	0.05907
LocPoly-p3-diff	1.96209	0.02835	5.06010	0.05959
xy	1.97966	0.02608	5.04778	0.05731
pwy	1.98021	0.02611	5.04479	0.05724
mcmb	1.97987	0.02604	5.04600	0.05751
wxy	1.98104	0.02606	5.04368	0.05716
riid	1.97719	0.02701	5.05784	0.05894
rnid	1.97719	0.02701	5.05784	0.05894
wiid	1.97719	0.02701	5.05784	0.05894
wker	1.97719	0.02701	5.05784	0.05894
wnid	1.97719	0.02701	5.05784	0.05894

Table E.13: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim \mathcal{N}(0, 16)$_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	93.80%	2.83644	0.00773	98.20%	4.91005	0.01336
LocPoly-p1-res	93.20%	2.84035	0.00784	98.40%	4.91767	0.01357
LocPoly-p1-diff	93.40%	2.83692	0.00776	98.40%	4.91209	0.01354
LocPoly-p3-diff	93.60%	3.89996	0.44810	98.20%	6.75096	0.77265
xy	94.20%	2.39966	0.01873	94.60%	5.14182	0.03541
pwy	94.00%	2.40367	0.01863	95.60%	5.15342	0.03538
mcmb	96.60%	2.62180	0.01687	94.80%	5.22008	0.03599
wxy	94.80%	2.39729	0.01856	95.00%	5.12560	0.03535
riid	89.40%	1.90432	0.01659	89.80%	4.03163	0.03338
rnid	90.00%	1.91831	0.01647	90.60%	4.14032	0.03414
wiid	97.80%	2.85712	0.01747	93.00%	4.94620	0.03025
wker	97.20%	2.98334	0.00634	98.80%	5.94844	0.01397
wnid	95.00%	2.42292	0.01300	94.60%	5.17704	0.01947

Table E.14: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim t(20)\_\text{nid}$

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00662	0.00667	4.98420	0.01460
LocPoly-p1-res	2.00628	0.00664	4.98505	0.01453
LocPoly-p1-diff	2.00567	0.00662	4.98510	0.01449
LocPoly-p3-diff	2.00692	0.00675	4.98468	0.01452
xy	2.00572	0.00662	4.98540	0.01408
pwy	2.00585	0.00662	4.98465	0.01407
mcmb	2.00545	0.00662	4.98559	0.01409
wxy	2.00618	0.00663	4.98483	0.01410
riid	2.00556	0.00686	4.98520	0.01452
rnid	2.00556	0.00686	4.98520	0.01452
wiid	2.00556	0.00686	4.98520	0.01452
wker	2.00556	0.00686	4.98520	0.01452
wnid	2.00556	0.00686	4.98520	0.01452

Table E.15: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim t(20)\_\text{nid}$

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.00%	0.72090	0.00198	94.60%	1.24903	0.00344
LocPoly-p1-res	98.60%	0.72069	0.00195	94.40%	1.24832	0.00345
LocPoly-p1-diff	99.00%	0.73206	0.01100	95.00%	1.26641	0.01898
LocPoly-p3-diff	98.40%	0.82242	0.03054	94.40%	1.42585	0.05366
xy	93.40%	0.60129	0.00452	94.80%	1.28837	0.00888
pwy	94.20%	0.60261	0.00457	95.20%	1.29032	0.00884
mcmb	96.20%	0.65736	0.00424	95.80%	1.31083	0.00916
wxy	94.40%	0.60130	0.00454	95.40%	1.28797	0.00899
riid	87.20%	0.48858	0.00425	88.00%	1.03025	0.00879
rnid	87.40%	0.49190	0.00423	89.00%	1.05890	0.00891
wiid	97.60%	0.71713	0.00446	94.00%	1.24148	0.00772
wker	98.80%	0.76071	0.00170	98.00%	1.51371	0.00379
wnid	96.00%	0.61111	0.00332	95.40%	1.30430	0.00495

Table E.16: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim t(10)\_\text{nid}$

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00074	0.00675	5.00005	0.01478
LocPoly-p1-res	2.00077	0.00675	4.99990	0.01479
LocPoly-p1-diff	2.00038	0.00678	4.99988	0.01482
LocPoly-p3-diff	2.00043	0.00680	5.00013	0.01480
xy	2.00179	0.00674	4.99984	0.01433
pwy	2.00186	0.00673	4.99918	0.01429
mcmb	2.00158	0.00673	4.99985	0.01430
wxy	2.00199	0.00674	4.99912	0.01429
riid	2.00217	0.00697	4.99996	0.01482
rnid	2.00217	0.00697	4.99996	0.01482
wiid	2.00217	0.00697	4.99996	0.01482
wker	2.00217	0.00697	4.99996	0.01482
wnid	2.00217	0.00697	4.99996	0.01482

Table E.17: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim t(10)\_\text{nid}$

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.20%	0.73502	0.00199	94.20%	1.27352	0.00344
LocPoly-p1-res	98.40%	0.73333	0.00194	94.20%	1.27039	0.00336
LocPoly-p1-diff	98.20%	0.75536	0.01648	94.00%	1.30830	0.02903
LocPoly-p3-diff	98.20%	0.80381	0.02564	94.00%	1.39194	0.04426
xy	94.40%	0.60564	0.00463	95.00%	1.30005	0.00898
pwy	95.20%	0.60823	0.00465	95.20%	1.30581	0.00900
mcmb	96.40%	0.66152	0.00417	95.20%	1.32074	0.00914
wxy	94.20%	0.60595	0.00459	94.60%	1.29808	0.00893
riid	88.20%	0.48562	0.00412	86.80%	1.03157	0.00852
rnid	88.40%	0.48961	0.00408	88.40%	1.06097	0.00871
wiid	97.20%	0.71629	0.00416	92.20%	1.24003	0.00721
wker	98.80%	0.77779	0.00181	97.00%	1.54913	0.00410
wnid	94.80%	0.61193	0.00312	94.80%	1.31830	0.00508

Table E.18: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

Method	mean variance of the standardised residuals	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	0.99372	0.00036
LocPoly-p1-ress	1.00274	0.00031
LocPoly-p1-diff	1.00582	0.00111
LocPoly-p3-diff	1.06658	0.02443
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	0.99505	0.00041
LocPoly-p1-ress	1.00173	0.00028
LocPoly-p1-diff	1.00516	0.00107
LocPoly-p3-diff	2.55498	0.66179
$e \sim t(20)$		
dglm.Tw.0	0.99430	0.00041
LocPoly-p1-ress	1.00298	0.00036
LocPoly-p1-diff	1.09221	0.08639
LocPoly-p3-diff	1.63566	0.18575
$e \sim t(10)$		
dglm.Tw.0	0.99490	0.00047
LocPoly-p1-ress	1.00365	0.00041
LocPoly-p1-diff	1.14390	0.10684
LocPoly-p3-diff	1.39148	0.13674

Table E.19: Average values of the variance of the standardised residuals when applying the conditional variance estimate obtained by Tweedie DGLM (link=log) model, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach and Local Polynomial of $p = 3$ difference-based approach.

Method	mean of the ratio of the determinants	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	0.96098	0.00932
LocPoly-p1-ress	0.95054	0.00905
LocPoly-p1-diff	0.95167	0.00876
LocPoly-p3-diff	1.11571	0.09200
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	0.93966	0.00829
LocPoly-p1-ress	0.94209	0.00815
LocPoly-p1-diff	0.93808	0.00811
LocPoly-p3-diff	20228.24441	13081.72999
$e \sim t(20)$		
dglm.Tw.0	0.97111	0.00875
LocPoly-p1-ress	0.96636	0.00840
LocPoly-p1-diff	7.84866	6.87935
LocPoly-p3-diff	90.10770	41.63468
$e \sim t(10)$		
dglm.Tw.0	1.01392	0.00979
LocPoly-p1-ress	1.00281	0.00957
LocPoly-p1-diff	30.64949	27.38842
LocPoly-p3-diff	60.50873	31.45844

Table E.20: Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed. The conditional variance is estimated by using Tweedie DGLM (link=log) model, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach and Local Polynomial of $p = 3$ difference-based approach.

Heteroscedastic, Location-Scale Shift Model With Linear Variance Function

The results of the problem of estimating the 50^{th} ($\tau = 0.5$) quantile function of the heteroscedastic, location-scale shift model (M2):

$$\text{Model 2} \quad : \quad y_i = 2 + 5x_i + \sigma(x_i)e_i$$

where $x \in [0, 1]$ and $x_i = i/n$ for $n = 500$ are presented i the following tables. The variance function is given by

$$\sigma^2(x_i) = \frac{1}{4}\exp(2x_i),$$

and the error, $\{e_i\}$, is iid from four different distributions: $\mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 16)$, $t(20)$ and $t(10)$.

$e \sim \mathcal{N}(0, 0.04)$ _nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00061	0.00065	4.99821	0.00151
LocPoly-p1-res	2.00061	0.00066	4.99824	0.00152
LocPoly-p1-diff	2.00061	0.00065	4.99821	0.00151
LocPoly-p3-diff	2.00057	0.00065	4.99833	0.00151
xy	2.00084	0.00063	4.99826	0.00145
pwy	2.00078	0.00063	4.99835	0.00145
mcmb	2.00085	0.00064	4.99831	0.00145
wxy	2.00085	0.00063	4.99815	0.00145
riid	2.00088	0.00066	4.99826	0.00151
rnid	2.00088	0.00066	4.99826	0.00151
wiid	2.00088	0.00066	4.99826	0.00151
wker	2.00088	0.00066	4.99826	0.00151
wnid	2.00088	0.00066	4.99826	0.00151

Table E.21: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim \mathcal{N}(0, 0.04)$_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.80%	0.06920	0.00019	91.80%	0.11988	0.00032
LocPoly-p1-res	98.40%	0.06930	0.00020	91.60%	0.11997	0.00034
LocPoly-p1-diff	98.60%	0.07138	0.00049	91.80%	0.12374	0.00084
LocPoly-p3-diff	98.80%	0.06958	0.00023	91.60%	0.12041	0.00040
xy	95.00%	0.05807	0.00045	95.80%	0.13508	0.00095
pwy	95.40%	0.05823	0.00044	95.80%	0.13557	0.00095
mcmb	97.40%	0.06502	0.00042	95.40%	0.13663	0.00096
wxy	96.00%	0.05796	0.00044	96.20%	0.13519	0.00095
riid	85.60%	0.04525	0.00038	85.40%	0.10302	0.00084
rnid	86.00%	0.04580	0.00037	86.80%	0.10811	0.00090
wiid	97.60%	0.06923	0.00043	91.60%	0.11985	0.00074
wker	99.40%	0.07277	0.00015	98.00%	0.15429	0.00038
wnid	93.00%	0.05469	0.00031	93.80%	0.12885	0.00048

Table E.22: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim \mathcal{N}(0, 16)$_nid

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.01226	0.01309	4.96424	0.03022
LocPoly-p1-res	2.01354	0.01322	4.96461	0.03032
LocPoly-p1-diff	2.01778	0.01413	4.96362	0.03048
LocPoly-p3-diff	2.01178	0.01314	4.96488	0.03021
xy	2.01689	0.01265	4.96509	0.02899
pwy	2.01566	0.01263	4.96691	0.02895
mcmb	2.01836	0.01267	4.96321	0.02907
wxy	2.01705	0.01262	4.96296	0.02892
riid	2.01761	0.01314	4.96511	0.03022
rnid	2.01761	0.01314	4.96511	0.03022
wiid	2.01761	0.01314	4.96511	0.03022
wker	2.01761	0.01314	4.96511	0.03022
wnid	2.01761	0.01314	4.96511	0.03022

Table E.23: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim \mathcal{N}(0, 16)$_nid

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.80%	1.38399	0.00376	91.80%	2.39749	0.00645
LocPoly-p1-res	98.40%	1.46747	0.05904	91.60%	2.53729	0.10005
LocPoly-p1-diff	98.60%	2.54761	0.21351	92.00%	4.44124	0.37488
LocPoly-p3-diff	98.80%	1.67348	0.09855	91.80%	2.89863	0.17128
xy	95.00%	1.16148	0.00894	95.80%	2.70156	0.01894
pwy	95.40%	1.16463	0.00889	95.80%	2.71132	0.01905
mcmb	97.20%	1.30035	0.00831	95.40%	2.73243	0.01917
wxy	96.00%	1.15902	0.00884	96.20%	2.70217	0.01886
riid	86.80%	0.90235	0.00726	87.00%	2.06041	0.01675
rnid	87.40%	0.91383	0.00722	88.80%	2.16523	0.01742
wiid	97.60%	1.38455	0.00856	91.60%	2.39692	0.01481
wker	99.40%	1.45547	0.00301	98.00%	3.08582	0.00762
wnid	93.00%	1.09388	0.00614	93.80%	2.57698	0.00952

Table E.24: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim t(20)\_\text{nid}$

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00208	0.00317	4.99207	0.00761
LocPoly-p1-res	2.00156	0.00318	4.99207	0.00760
LocPoly-p1-diff	2.00096	0.00320	4.99150	0.00761
LocPoly-p3-diff	2.00180	0.00315	4.99184	0.00759
xy	2.00203	0.00303	4.99150	0.00725
pwy	2.00225	0.00304	4.99089	0.00727
mcmb	2.00172	0.00305	4.99184	0.00730
wxy	2.00202	0.00303	4.99186	0.00725
riid	2.00210	0.00320	4.99197	0.00760
rnid	2.00210	0.00320	4.99197	0.00760
wiid	2.00210	0.00320	4.99197	0.00760
wker	2.00210	0.00320	4.99197	0.00760
wnid	2.00210	0.00320	4.99197	0.00760

Table E.25: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim t(20)\_\text{nid}$

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.80%	0.35414	0.00095	92.40%	0.61317	0.00168
LocPoly-p1-res	99.00%	0.36834	0.00571	92.40%	0.63826	0.00978
LocPoly-p1-diff	99.00%	0.43959	0.01648	93.20%	0.76246	0.02876
LocPoly-p3-diff	99.00%	0.39419	0.01178	92.80%	0.68318	0.02041
xy	96.80%	0.29472	0.00226	95.60%	0.68846	0.00499
pwy	96.60%	0.29502	0.00226	95.60%	0.68868	0.00491
mcmb	97.80%	0.32974	0.00212	96.20%	0.69376	0.00496
wxy	96.40%	0.29383	0.00224	95.60%	0.68687	0.00496
riid	86.20%	0.22943	0.00199	85.40%	0.52167	0.00412
rnid	87.20%	0.23250	0.00197	87.40%	0.54616	0.00428
wiid	98.20%	0.35134	0.00211	91.60%	0.60824	0.00366
wker	98.80%	0.37269	0.00082	97.40%	0.79166	0.00204
wnid	95.60%	0.27836	0.00157	94.20%	0.65465	0.00247

Table E.26: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

$e \sim t(10)\_\text{nid}$

	β_0		β_1	
Method	b_0	SE	b_1	SE
dglm.Tw.0	2.00013	0.00329	5.00046	0.00777
LocPoly-p1-res	1.99999	0.00329	5.00062	0.00777
LocPoly-p1-diff	2.00099	0.00333	4.99972	0.00782
LocPoly-p3-diff	1.99993	0.00329	5.00054	0.00779
xy	2.00082	0.00326	4.99996	0.00754
pwy	2.00086	0.00326	4.99962	0.00752
mcmb	2.00063	0.00325	5.00026	0.00749
wxy	2.00094	0.00326	4.99956	0.00752
riid	2.00101	0.00336	5.00045	0.00779
rnid	2.00101	0.00336	5.00045	0.00779
wiid	2.00101	0.00336	5.00045	0.00779
wker	2.00101	0.00336	5.00045	0.00779
wnid	2.00101	0.00336	5.00045	0.00779

Table E.27: Parameter estimates for Model 2 for $\tau = 0.5$.

$e \sim t(10)\_\text{nid}$

	β_0			β_1		
Method	C	L	SE	C	L	SE
dglm.Tw.0	98.60%	0.35891	0.00098	91.40%	0.62176	0.00169
LocPoly-p1-res	98.60%	0.38070	0.00773	92.00%	0.65967	0.01347
LocPoly-p1-diff	98.60%	0.47330	0.01990	92.60%	0.82008	0.03462
LocPoly-p3-diff	99.00%	0.39239	0.00853	91.80%	0.67942	0.01465
xy	94.20%	0.29332	0.00221	95.20%	0.68270	0.00474
pwy	94.80%	0.29462	0.00222	95.40%	0.68552	0.00475
mcmb	96.40%	0.32898	0.00205	95.00%	0.69034	0.00481
wxy	94.60%	0.29343	0.00218	95.20%	0.68161	0.00471
riid	91.40%	0.23257	0.00188	87.60%	0.52876	0.00409
rnid	92.00%	0.23584	0.00186	89.40%	0.55616	0.00428
wiid	97.00%	0.35051	0.00202	90.60%	0.60680	0.00349
wker	98.80%	0.38041	0.00090	97.00%	0.80559	0.00227
wnid	94.00%	0.27974	0.00152	93.60%	0.65966	0.00254

Table E.28: Coverage probabilities (C) and average lengths of the 95% confidence intervals (L) for each parameter.

Method	mean variance of the standardised residuals	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	1.00610	0.00148
LocPoly-p1-ress	1.01244	0.00797
LocPoly-p1-diff	1.23489	0.04364
LocPoly-p3-diff	1.04227	0.01239
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	1.00610	0.00148
LocPoly-p1-ress	1.28242	0.20033
LocPoly-p1-diff	6.32080	1.02245
LocPoly-p3-diff	1.94073	0.32709
$e \sim t(20)$		
dglm.Tw.0	1.01543	0.00630
LocPoly-p1-ress	1.20309	0.07682
LocPoly-p1-diff	2.38898	0.26917
LocPoly-p3-diff	1.50331	0.13958
dglm.Tw.0	1.01099	0.00181
LocPoly-p1-ress	1.31455	0.11310
LocPoly-p1-diff	2.77639	0.32498
LocPoly-p3-diff	1.40826	0.09733

Table E.29: Average values of the variance of the standardised residuals when applying the conditional variance estimate obtained by Tweedie DGLM (link=log) model, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach and Local Polynomial of $p = 3$ difference-based approach.

Method	mean of the ratio of the determinants	SE
$e \sim \mathcal{N}(0, 0.04)$		
dglm.Tw.0	0.96170	0.00927
LocPoly-p1-ress	0.96604	0.01164
LocPoly-p1-diff	1.28890	0.07544
LocPoly-p3-diff	0.99154	0.01863
$e \sim \mathcal{N}(0, 16)$		
dglm.Tw.0	0.96176	0.00927
LocPoly-p1-ress	248.72751	175.06245
LocPoly-p1-diff	3810.69847	865.57012
LocPoly-p3-diff	629.06375	340.73347
$e \sim t(20)$		
dglm.Tw.0	0.98585	0.00937
LocPoly-p1-ress	4.71893	1.62193
LocPoly-p1-diff	61.84624	15.09157
LocPoly-p3-diff	34.94683	13.07994
$e \sim t(10)$		
dglm.Tw.0	1.01329	0.00997
LocPoly-p1-ress	13.42218	5.31459
LocPoly-p1-diff	118.85786	29.69786
LocPoly-p3-diff	10.50885	3.38507

Table E.30: Mean of the ratio of the determinants of the estimated covariance matrices of $\boldsymbol{b}$ when applying kernel smoothing bootstrapping adjusted to have first and the second moment the same as the data from which it is constructed. The conditional variance is estimated by using Tweedie DGLM (link=log) model, Local Polynomial of $p = 1$ residual-based approach, Local Polynomial of $p = 1$ difference-based approach and Local Polynomial of $p = 3$ difference-based approach.